INSTITUT FÜR INFORMATIK
Datenbanken und Informationssysteme

Universitätsstr. 1 D–40225 Düsseldorf

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Analyse von Wiktionary als Wortnetz

Jan Oberpichler

Bachelorarbeit

Beginn der Arbeit:	18. Mai 2017
Abgabe der Arbeit:	18. August 2017
Gutachter:	Prof. Dr. Stefan Conrad
	Prof. Dr. Michael Schöttner

Erklärung

Hiermit versichere ich, dass ich diese Bachelorarbeit selbstständig verfasst habe. Ich habe dazu keine anderen als die angegebenen Quellen und Hilfsmittel verwendet.

Düsseldorf, den 18. August 2017

_____

Jan Oberpichler

Zusammenfassung

In dieser Arbeit wird untersucht, wie weit sich das deutsche Wiktionary - ein von freiwilligen Nutzern aufgebautes Onlinelexikon - als Wortnetz eignet. Hierzu wird zunächst aus Wiktionary ein Wortnetz konstruiert und anschließend GermaNet, ein bereits seit 1998 bestehendes und fortwährend ausgebautes Wortnetz für die deutsche Sprache, zum Vergleich herangezogen.
Zunächst wird das Extrahieren der für das zu konstruierende Wortnetz relevanten Daten aus dem frei verfügbaren, monatlichen XML-Dump Wiktionarys beschrieben. Anschließend werden unterschiedliche Qualitäten und Nachteile über verschiedene Vergleichskriterien identifiziert, die vor allem den Umfang und die Vernetzung des konstruierten Wiktionary-Wortnetzes gegenüber GermaNet abtasten.

Es zeigt sich, dass Wiktionary trotz seines großen Wachstums seit dem Start im Mai 2004 GermaNet an Größe und Umfang unterliegt. Für die meisten Vergleiche ist dies ein signifikanter Faktor, der das Wiktionary-Wortnetz gegenüber GermaNet benachteiligt. Außerdem ist die Struktur eines Wiktionary-Artikels nicht gänzlich ideal für die Konstruktion eines Wortnetzes, wodurch das resultierende Netz Informationen gröber miteinander verknüpft als GermaNet.

Grundsätzlich stellt sich jedoch heraus, dass bereits im aktuellen Zustand Wiktionary ein solides Wortnetz darstellt, welches sich über verschiedene Qualitäten einzigartig hervorhebt. Besonders die frequente Aufnahme moderner Begriffe ist für Wiktionary ein Alleinstellungsmerkmal. Zudem wurde eine besonders hohe Abdeckung alemannischer Begriffe festgestellt.
Wiktionary-Einträge beschreiben ein Wort und seine Zusammenhänge häufig sogar ausführlicher als GermaNet. Die meisten Begriffe im Wiktionary-Wortnetz sind granularer in verschiedene Bedeutungen aufgeteilt. Der Vergleich über ein word embedding zeigt darüber hinaus, dass syntaktische und semantische Zusammenhänge im Wiktionary-Wortnetz besser kodiert werden als in GermaNet.

Im Fazit wird letztlich eine Übersicht über alle Vergleichskriterien gegeben und die Eigenschaften des Wiktionary-Wortnetzes evaluiert. Das deutschsprachige Wiktionary zeigt sich als eine Baustelle, die trotz großen Arbeitsbedarfs bereits ein gutes Gerüst für ein Wortnetz bildet.

Inhaltsverzeichnis

1 Einleitung **1**

1.1 Ziel 1

1.2 Präliminarien 1

1.3 Wortnetze und ihre Anwendungsgebiete 2

2 GermaNet und Wiktionary **3**

2.1 GermaNet 3

2.2 Wiktionary als Wortnetz 4

3 Vergleich **8**

3.1 Deskriptive Statistiken 8

3.2 Qualität der Vernetzung 11

3.3 Abdeckung 14

3.4 Vergleich über *word embeddings* 21

3.5 Vergleich über eine Umfrage 26

4 Fazit **29**

4.1 Resultate und Evaluation 29

4.2 Zukünftige Arbeit 31

A Anhang **32**

Literatur **35**

Abbildungsverzeichnis **37**

Tabellenverzeichnis **37**

1 Einleitung

In diesem einleitenden Kapitel wird zunächst das Ziel der Arbeit genau formuliert. Anschließend werden grundlegende Begrifflichkeiten geklärt und die Anwendungsgebiete von Wortnetzen überblickt, um eine Grundlage für den Vergleich zu schaffen.

1.1 Ziel

Ziel der Arbeit ist der Vergleich zwischen dem als Wortnetz interpretierten Wiktionary[1], dem Wiki-basiertem freien Wörterbuch für die deutsche Sprache, und dem Wortnetz GermaNet 9.0[2] der Universität Tübingen [HF97]. Entscheidender Unterschied ist hierbei, dass Wiktionary als kollaborative Wissensbasis von Freiwilligen aufgebaut und aktualisiert wird während GermaNet von Linguisten nach bestimmten Richtlinien erstellt wird (linguistische Wissensbasis).
Aufgabe ist es also zunächst, die im Wiktionary eingetragenen Daten zu einem Wortnetz zu verarbeiten, was über Parsen eines öffentlich zugänglichen XML-Dumps zu realisieren ist. Das resultierende Wiktionary-Wortnetz wird dann mit GermaNet verglichen. Dabei sollen konkrete Stärken und Schwächen der beiden Wortnetze im Vergleich miteinander aufgezeigt werden und somit idealerweise auch ein Überblick darüber verschafft werden, in welchen Anwendungsbereichen die Wortnetze jeweils besonders nutzbringend einsetzbar sind. Dazu werden relevante Vergleichskriterien herangezogen und für beide Wortnetze evaluiert.

1.2 Präliminarien

Wortnetze sind lexikalisch-semantische, elektronische Ressourcen für bedeutungstragende Beziehungen einer oder mehrerer Sprachen [CEE+09]. Als grundlegendes Konzept für diese Arbeit werden beide Wortnetze in Kapitel 2 näher betrachtet. Zudem wird eine Übersicht zu Wortnetzen und ihren Anwendungsgebieten im nächsten Abschnitt gegeben.

Da die Diskussion über Wortnetze und Wörter im Allgemeinen zu teilweise verwirrenden terminologischen Bedeutungsunklarheiten führen kann, werden zunächst die wichtigsten Termini und deren Gebrauch in dieser Arbeit geklärt:
Als **Lexem** wird die atomarste Bedeutungseinheit bezeichnet. In dieser Arbeit umfasst dies immer eine Schreibweise und eine eindeutig zugehörige Bedeutung. Die Lexeme „Fahrbahn“ (Fahrfläche einer Straße) und „Spur“ (ebenfalls Fahrfläche einer Straße) unterscheiden sich lediglich in ihrer Schreibweise, während sich das Lexem „Spur“ (hinweisgebende Hinterlassenschaft) nur durch seine Bedeutung zu letzterem Lexemen unterscheidet.
Werden alle Bedeutungen einer Schreibweise zusammengefasst so wird dies als **Term** bezeichnet. Die „Spur“ mit all seinen Bedeutungen wäre ein solcher Term. Dies kommt dem allgemeinen Verständnis eines Wortes also näher, wodurch sich *Wort* oder *Begriff*

[1] `https://de.wiktionary.org/wiki/Wiktionary:Hauptseite`

[2] GermaNet 9.0 wurde im April 2014 veröffentlicht. Die aktuellste GermaNet Version ist Version 12.0, die aber zum Zeitpunkt dieser Arbeit leider noch nicht zur Verfügung stand.

ebenfalls gleichbedeutend eignen würden. Um jedoch konkret auf die alle Bedeutungen umfassende Schreibweise zu verweisen wird explizit der Begriff Term genutzt. Terme, die von mehr als einem Lexem beschrieben werden, heißen *polysem*.
Darüber hinaus werden mit **Bedeutungen** die einem Term zugehörigen Lexeme bezeichnet. Der polyseme Term „Viertel“ hat also zwei Bedeutungen, das mathematische Viertel und das Viertel als eine Wohngegend, welche beide wiederum ein Lexem mit der Schreibweise „Viertel“ sind.

1.3 Wortnetze und ihre Anwendungsgebiete

Wortnetze sind eine elektronische Datenbasis für Relationen zwischen Synsets. Als **Synset** wird eine Menge von Lexemen bezeichnet, die eine bestimmte, in jedem Lexem der Menge gleichermaßen getragene Bedeutung repräsentiert. So bezeichnet zum Beispiel in der Menge {Bank, Kasse, Geldinstitut, Geldhaus} jedes Lexem eine „Bank“ als Geldinstitut. Ein Lexem ist einem Synset durch eine ID eindeutig zugewiesen. Da der Term „Bank“ nicht nur das Geldinstitut, sondern auch eine Sitzgelegenheit für mehrere Personen bezeichnet, existiert außerdem ein weiteres Lexem „Bank“ in dem Synset {Sitzbank, Bank} mit einer anderen ID. Dadurch ist die Polysemie des Terms „Bank“ gegeben, während alle Schreibweisen der Bedeutungen in den Synsets zusammengefasst sind.
Die Synsets sind über **Relationen** verbunden. Sie bilden somit einen Graphen in dem Synsets die Knoten und Relationen die Kanten darstellen. Ob eine Kante gerichtet oder ungerichtet ist hängt dabei vom Relationstyp ab. Welche relativen Zusammenhänge zwischen den Synsets repräsentiert werden hängt vom jeweiligen Wortnetz ab. Diese Arbeit beschränkt sich auf auf die Synonym-, Hyponym-, Hyperonym- und Antonymrelationen.

Da sich die Anordnung von über Relationen verbundenen Synsets als Graph interpretieren lässt, sind nicht zuletzt graphentheoretische Algorithmen für die Mobilisierung von Wortnetzen von Bedeutung. Aber auch für *machine learning* Algorithmen sind Wortnetze von Interesse. So finden sich Anwendungen beispielsweise im Bereich der *word-sense disambiguation* [BP02], wobei versucht wird, die korrekte Bedeutung eines ambiguen Wortes in einem Satz zu bestimmen, oder der automatischen Textklassifikation [SM98] und Textzusammenfassung [BE99], bei der ein ganzer Text einer oder mehreren Klassen zugeordnet beziehungsweise zusammengefasst werden soll. Auch bei der maschinellen Übersetzung [KL94] werden Wortnetze eingesetzt. Besonders interessant sind Wortnetze außerdem für die Berechnung der semantischen Nähe zweier Wörter [BH06], die eine numerische Beschreibung der Sinnverwandheit zweier Wörter darzustellen versucht.

2 GermaNet und Wiktionary

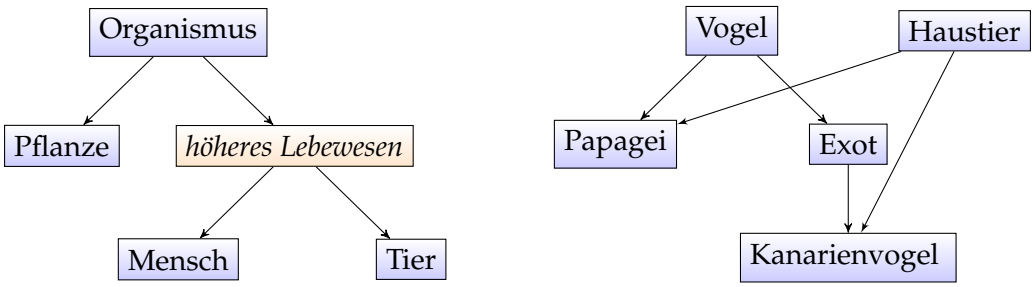
Dieses Kapitel widmet sich zunächst dem Wortnetz GermaNet, seiner Struktur und für den Vergleich zu beachtende besondere Charakteristiken. Darauf folgt eine schematische Beschreibung des Wiktionary Parsers, der die Daten aus Wiktionary in eine geeignete Wortnetzarchitektur überführt, sowie die Diskussion des Aufbaus und der Struktur des resultierenden Wiktionary-Wortnetzes.

2.1 GermaNet

Das lexikalisch-semantische Wortnetz GermaNet orientiert sich strukturell grundlegend an dem Princeton WordNet [Mil95]. Dementsprechend ist GermaNet aus den zwei in 1.3 beschriebenen Konzepten aufgebaut: Synsets und Relationen.
In GermaNet kodierte Relationen sind Synonymie, Antonymie, Hyperonymie/Hyponymie, Meronymie/Holonymie, Kausalität, Assoziativität, Petronymie und Partizipation. Für die weiteren Betrachtungen sind aber lediglich Synonymie, Antonymie und Hypero-/Hyponymie von Interesse.

Ein Synset beinhaltet ein oder mehrere *LexUnits* und gehört zu einer Wortart. *LexUnits* sind hierbei das Äquivalent zu Lexemen. Die ID der *LexUnits* wird, beginnend mit 1, für jede Bedeutung einer Schreibweise hochgezählt. So ist beispielsweise „Haus(4)“ das Lexem „Haus“ als das „zu einem bestimmten Zweck gebaute Gebäude“ aus dem Synset {Haus(4)}, welches keine weiteren *LexUnits* umfasst. Die *LexUnit* „Haus(4)“ ist dabei eindeutig diesem Synset zugewiesen und kann in keinem anderen Synset vorkommen.

GermaNet macht darüber hinaus Gebrauch von **künstlichen Konzepten**. Sie dienen der besseren Strukturierung von Hypero-/Hyponymbeziehungen, indem sie lexikographische Lücken füllen.
Ein künstliches Konzept ist ein Synset mit einem artifiziellen Lexem, einem *Konzept*. Synsets, die ein künstliches Konzept beinhalten, sind immer einelementig und besonders gekennzeichnet. Sie bilden also lexikalische Bedeutungen ab, wie sie im Deutschen nicht existieren, aber zur sinnvollen Strukturierung vorteilhaft sind (Beispiel: 1a). Ein künstliches Konzept kann durch ein Lexem einer anderen Sprache beschrieben sein, was aber nicht zwingend der Fall sein muss.



(a) Künstliches Konzept (b) Kreuzklassifizierung

Abbildung 1: GermaNet Konzepte

Im Gegensatz zu WordNet lassen sich außerdem Synsets kreuzklassifizieren, also kann jedes Synset mehrere hierarchische Relationen besitzen (Beispiel: 1b).
Überdies machen etwa 2,39% der *LexUnits* Mehrwortlexeme aus während Eigennamen außer für einige geografische Lexeme wie Städtenamen nicht repräsentiert sind [HF97].

2.2 Wiktionary als Wortnetz

Wiktionary ist ein Wikimedia-Projekt [3] mit dem Ziel, ein frei zugängliches Wörterbuch zu schaffen. Dabei ist es grundsätzlich jedem möglich, Einträge im Wiktionary zu verfassen, anzupassen oder zu erweitern. Ein **Eintrag** ist hierbei die für ein Wort erstellte Seite im Wiktionary.
Die Autoren sind auch für die Prüfung der Einträge zuständig. Beides erfolgt nach gewissen Richtlinien[4], wobei die Einträge selbst einer festgelegten Struktur unterliegen[5]. Diese Struktur eignet sich prinzipiell bereits dazu, sie als Wortnetz zu interpretieren.

Das Wiktionary wird monatlich in einem XML-Dump gesichert, auf den frei zugegriffen werden kann[6]. Es wird also ein Parser genutzt, um die relevanten Informationen jedes Eintrags zu extrahieren und in einer geeigneten Datenstruktur zu speichern. Diese Datenstruktur wird das Wiktionary-Wortnetz darstellen.

Da Wiktionary Einträge stets alle Bedeutungen, Flexionen und Ursprünge eines Wortes abzudecken versuchen und sich dies auch später in der Struktur des Wiktionary-Wortnetzes widerspiegeln wird, kommt hier der in Kapitel 1.2 eingeführte Begriff Term besonders zum tragen. Jeder Wiktionary-Eintrag repräsentiert einen Term mit all seinen Bedeutungen, also die Lexeme gleicher Schreibweise.

2.2.1 Struktur eines Wiktionary Artikels

Quellcode 1 stellt eine gekürzte und angepasste Version des Wiktionary Eintrags zu „Seite“[7] in Wiki-Syntax dar, an dem sich die Struktur gut erkennen lässt: Jeder Eintrag beginnt mit einer *Level-2-Überschrift*, die den Term und die (Ursprungs-)Sprache kodiert. Level-2-Überschriften beginnen und enden mit einem doppelten Gleichheitszeichen. Sie enhalten Informationen über die Wortart, die Sprache im Bezug auf das Vorkommen und, im Falle von Substantiven, auch den Genus. Es existieren beispielsweise zwei Einträge für American Football: „American football“[8] als substantive Wortverbindung im Englischen und „American Football“[9] als substantiver, maskuliner Eigenname im Deutschen. Dies ist nicht zu verwechseln mit dem Eintrag für „American football“[10] im englischen Wiktionary, welches komplett unabhängig vom Deutschen existiert.
Dementsprechend beginnen und enden *Level-3-Überschriften* mit dreifachen

[3] https://www.wikimedia.de/wiki/Hauptseite
[4] https://de.wiktionary.org/wiki/Hilfe:Allgemeines_zu_Eintr%C3%A4gen
[5] https://de.wiktionary.org/wiki/Hilfe:Eintrag
[6] https://dumps.wikimedia.org/dewiktionary/
[7] https://de.wiktionary.org/wiki/Seite
[8] https://de.wiktionary.org/wiki/American_football
[9] https://de.wiktionary.org/wiki/American_Football
[10] https://en.wiktionary.org/wiki/American_football

Gleichheitszeichen - auf jede Level-2-Überschrift folgt mindestens eine Level-3-Überschrift. Jede Level-3-Überschrift kodiert verschiedene Genera, Flexionen, oder Ursprünge eines Terms. Beispielsweise gäbe es für den Term „Bock“ zwei Level-3-Überschriften, da das Lexem „Bock“ im Neutrum die Bedeutung „Bier mit einem Stammwürzegehalt über 16%“ und im Maskulinum die Bedeutung „ein vierbeiniges oder klobiges Gestell“ trägt. Für das Wiktionary-Wortnetz werden die Level-3-Überschriften zusammengefasst und auf den zugrundeliegenden Term reduziert, so dass ein Term-Knoten alle Genera der Schreibweise umfasst (sowie deren Bedeutungen).

Jeder Level-3-Abschnitt enthält zudem kategorisierte Informationen zu dem behandelten Term. Eine Kategorie-Überschrift ist in doppelten geschweiften Klammern notiert. Abseits der für das zu konstruierende Wortnetz relevanten Informationen können noch einige andere Daten eingetragen sein. Dazu gehören beispielsweise Flexionstabellen (siehe Zeile 4-6), Worttrennung, Abkürzungen, Herkunft, sinnverwandte Wörter, Beispiele und mehr. Die für das Wortnetz wichtigste Kategorie ist dabei die Kategorie „Bedeutungen“, die Informationen über die Anzahl der Bedeutungen und eine Beschreibung eben jener enthält.
Einer Bedeutung ist jeweils eine eindeutige Zahl zugewiesen (Bsp. Zeile 12). Für kleinere Bedeutungsunterschiede kann eine Bedeutung auch noch weiter in Abschnitte wie a und b unterteilt werden. Die vorangestellten Doppelpunkte kodieren dabei das Level der Aufteilung. Für das Wortnetz werden jedoch die Bedeutungsbeschreibungen auf dem allgemeinsten Level zusammengefasst, sodass unabhängig von der Aufteilungsgranularität nur eine Bedeutung (beziehungsweise ein Lexem) interpretiert wird.

Unter den Kategorien „Synonyme“, „Gegenwörter“, „Oberbegriffe“ und „Unterbegriffe“ sind entsprechende, in Relation zusammenhängende Terme eingetragen. Jede Zeile in einer diesen Kategorien wird mit den zugehörigen Bedeutungen kodiert. Zum Beispiel sind in Zeile 27 die Terme „Profil“ und „Rand“ synonym jeweils zu den Bedeutungen 1 (a und b) und 2 von „Seite“. Somit lässt sich schließlich ein Wortnetz extrahieren, das Term-Knoten mit diesen Kanten verknüpft.

2.2.2 Struktur des Wiktionary-Wortnetzes

Anders als in GermaNet, welches jedes Lexem mit einer eindeutigen Bedeutung in den Knoten repräsentiert (Synsets), bildet im Wiktionary-Wortnetz ein ganzer Term einen Knoten. Zwar lässt sich jeder Term auf seine Lexeme herunterbrechen, jedoch verlieren die Kanten zwischen den Termen damit an Information. Während in GermaNet einem Lexem eindeutig Hypero- und Hyponyme zugewiesen werden können (beispielsweise einer „Bank“ (Geldinstitut) der Oberbegriff „Kapitalgeber“ und der Unterbegriff „Direktbank“) ist im Wiktionary unklar, zu welcher Bedeutung von „Linie“ das Lexem „Seite“ (als geometrische Grenzlinie eines Vielecks) Unterbegriff ist, wenn Linie als Oberbegriff zu Seite eingetragen ist.

Während sich im Wiktionary-Wortnetz keine künstlichen Konzepte (Abbildung 1a) finden, sind Terme beziehungsweise deren Lexeme ebenso wie in GermaNet kreuzklassifiziert (Abbildung 1b). Damit zeichnet sich ein weiterer signifikanter Unterschied zu GermaNet ab. So findet sich beispielsweise keine Hyponymkette von „Organismus“ zu „Mensch“, da das künstliche Konzept „*höheres Lebewesen*“ fehlt.

```
== Seite ({{Sprache|Deutsch}}) ==
=== {{Wortart|Substantiv|Deutsch}}, {{f}} ===

{{Deutsch Substantiv Übersicht
*...*
}}

{{Worttrennung}}
:Sei·te, {{Pl.}} Sei·ten

{{Bedeutungen}}
:[1] in einer bestimmten Richtung liegende Begrenzungsfläche
::[a] eines Gegenstandes nach außen
::[b] eines Raumes
:[2] {{K|Geometrie}} Grenzlinie eines Vieleckes
:[3] [[Blickwinkel]], [[Perspektive]]
:[4] {{K|Internet}} ''kurz für'' [[Internetseite]]

{{Abkürzungen}}
*...*

{{Herkunft}}
*...*

{{Synonyme}}
:[1] [[Teil]], [[Hälfte]], [[Flanke]], [[Richtung]]
:[1, 2] [[Profil]], [[Rand]],
:[3] [[Blickwinkel]], [[Warte]]

{{Gegenwörter}}
:[1] [[Achse]], [[Länge]], [[Mitte]], [[Zentrum]]

{{Oberbegriffe}}
:[1] [[Grenzfläche]]
:[2] [[Linie]]
:[4] [[Auftritt]]

{{Unterbegriffe}}
:[1] [[Backbordseite]], [[Breitseite]], [[Hinterfront]]
:[4] [[Homepage]] [[Anmeldeseite]], [[Hauptseite]]

*...*
```

Quellcode 1: Exemplarischer Wiktionaryartikel

Ein letzter entscheidender Unterschied ist die Synonymie. In GermaNet sind zueinander synonyme Lexeme in einem Synset zusammengefasst. Im Wiktionary-Wortnetz jedoch wird Synonymität durch eine Relationskante wie für Hypero-, Hypo- und Antonymie kodiert. Diese Kanten leiden unter dem bereits beschriebenen Präzisionsverlust durch Term-Knoten. Folglich fehlt dem Wiktionary-Wortnetz das Synset-Konzept gänzlich. Jede Bedeutung eines Term-Knoten lässt sich jedoch für den Vergleich als einelementiges Synset interpretieren.
Hierbei ist der Unterschied in der Datenstruktur beider Wortnetze wichtig: Während in GermaNet zueinander synonyme *Lexeme* zusammenhängend hinterlegt sind, befindet sich im Wiktionary-Wortnetz lediglich ein Verweis, eine Kante zu den synonymen *Termen* eines Lexems.

2.2.3 Probleme mit Irregularitäten und Unvollständigkeit

Das Wiktionary wird von vielen, nicht professionellen Nutzern erweitert, die versuchen, den Richtlinien bestmöglich zu folgen. Dementsprechend sind Wiktionary Einträge unvollständig, nicht einheitlich und können diverse Fehler enthalten. Für das resultierende Wortnetz ist besonders die Inkonsistenz der Relationen problematisch. Ist „Kammer“ beispielsweise als Synonym zu „Haus“ eingetragen, ist der umgekehrte Fall nicht zwingend gegeben - obwohl Synonymie dies impliziert. Gleiches gilt für Antonyme sowie Ober- und Unterbegriffe, die zueinander symmetrisch sein sollten.

Um das Ausmaß dieser Problematik besser untersuchen zu können werden zwei verschiedene Wiktionary-Wortnetze erstellen. Das Erste, **native Wiktionary-Wortnetz**, wird, wie in diesem Kapitel beschrieben, die Daten aus dem Wiktionary möglichst exakt übernehmen - also eben auch alle Irregularitäten in den Relationen.
Das Zweite, **symmetrische Wiktionary-Wortnetz**, wird so angepasst, dass die Relationen wieder symmetrisch zueinander sind. Ist also beispielsweise „Ort“ ein Oberbegriff zu „Haus“, ist „Haus“ auch ein Unterbegriff zu „Ort“.

Das Wiktionary-Wortnetz symmetrisch zu konstruieren bringt allerdings ein anderes Problem mit sich. Da unklar ist zu welcher Bedeutung eines Terms ein (beispielsweise) eingetragenes Synonym gehört wird der asymmetrische Term in jede Bedeutung des über eine Relation verknüpften Begriffs eintragen.
Ist also für „Fenster“ (beziehungsweise für eine bestimmte Bedeutung des Terms „Fenster“) der Term „Mauer“ asymmetrisch als Gegenwort eingetragen, wird für jede Bedeutung, die für den Term „Mauer“ definiert ist, „Fenster“ als Gegenwort eingetragen. Allerdings ungeachtet dessen, ob die einzelnen Bedeutungen jeweils auch tatsächlich antonym zu „Fenster“ sind.

3 Vergleich

Im Folgenden werden die beiden Wortnetze anhand von verschiedenen Kriterien verglichen und die Ergebnisse evaluieren.

3.1 Deskriptive Statistiken

Zunächst werden einige Statistiken zu den beiden Wortnetzen betrachtet, die einen groben Überblick zum Umfang und dem Grad der Vernetzung liefern.

3.1.1 Quantität

In Tabelle 1 findet sich ein Größenvergleich der beiden Netze. Hierbei wurden jeweils die Anzahl der Synsets im Wiktionary-Netz und GermaNet gezählt, wobei, wie bereits in Kapitel 2.2.2 beschrieben, im Wiktionary-Wortnetz jede Bedeutung eines Terms als einelementiges Synset interpretiert wird.

	Wiktionary	GermaNet
Substantive	*93 229*	*71 575*
Verben	*14 607*	*11 026*
Adjektive	*13 384*	*10 645*
Insgesamt	*121 220*	*93 246*

Tabelle 1: Anzahl an Synsets in Wiktionary und GermaNet

Der direkte Vergleich zeigt, dass das Wiktionary-Wortnetz um etwa 30% mehr Synsets kodiert als GermaNet, wobei der größte Unterschied, mit etwa 32% mehr Synsets im Wiktionary-Wortnetz, bei den Verben liegt. Nach den Angaben auf der GermaNet Website[11] ist dieser Unterschied mit der neusten GermaNet Version (Mai 2017) bereits nichtig, da nun bereits *120 032* Synsets eingetragen sind - was etwa der Größe des Wiktionary-Wortnetzes entspricht. Es ist jedoch zu beachten, dass hier nur die *Anzahl der Synsets* untersucht wurde, welche im Wiktionary-Wortnetz zusätzlich dadurch geschönt ist, da kein Synset mehrere Lexeme auf einmal zusammenfasst. Entscheidender wird die Anzahl der Terme, also informell gesprochen die Anzahl der nachschlagbaren Wörter, für die in Kapitel 3.3.1 festgestellt wird, dass GermaNet tatsächlich mehr Terme umfasst als Wiktionary. Wieso trotzdem mehr Synsets im Wiktionary-Wortnetz kodiert sind wird in Kapitel 3.1.3 untersucht.

Betrachtet man das Wachstum der beiden Wortnetze in den letzten Jahren, ist jedoch davon auszugehen, dass das Wiktionary-Wortnetz GermaNet in Zukunft an Größe übertreffen wird. Dies ist einer der großen Vorteile von kollaborativen Wissensbasen - aktuell jedoch übertrifft GermaNet in seinem Umfang noch das Wiktionary-Wortnetz.

[11] `http://www.sfs.uni-tuebingen.de/GermaNet/`

3.1.2 Relationen

Die folgenden Tabellen 2, 3 und 4 enthalten detaillierte Statistiken zu den Relationen im Wiktionary-Wortnetz und GermaNet. Hierbei wird zusätzlich noch zwischen der „nativ" übernommenen Version und der in Abschnitt 2.2.3 beschriebenen symmetrischen Version unterschieden.
Da im Wiktionary-Wortnetz jede Relation über eine Kante kodiert ist, werden hier einfach die von jeder Bedeutung ausgehenden Kanten pro Kantentyp gezählt. In GermaNet sind Hypo-, Hypero- und Antonymie ebenfalls über Kanten realisiert, für Synonyme wurden jedoch alle Lexeme gezählt, die zusätzlich zu mindestens einem anderen Lexem in einem Synset vorkommen (und somit synonym sind).
Ebenfalls zu beachten ist, dass jede Relation als *gerichtete* Kante gezählt wird. Antonym- und Synonymrelationen werden in GermaNet also jeweils doppelt gezählt (hin und zurück), während im Wiktionary-Wortnetz auch der einseitige (asymmetrische) Fall auftreten kann und somit nur eine einzelne Relationskante gezählt wird. In einem Synset {Baumkrone, Wipfel, Baumwipfel} finden sich also sechs Synonymrelationen, während in einem Synset {Lehrmittel} keine Synonymrelationen gezählt werden.
Die Unterscheidung von Hypo- und Hyperonymrelationen ist überhaupt nur deshalb nötig, da die Relationen wie in Kapitel 2.2.3 beschrieben in Wiktionary nicht zwingend symmetrisch eingetragen sind.

Relation	Substantive	Verben	Adjektive	Insgesamt
Σ Hyponyme (insgesamt)	*75 541*	*11 312*	*10 859*	*97 712*
⌀ Hyponyme (pro Bedeutung)	*1.055410*	*1.025938*	*1.020103*	*1.047894*
Σ Hyperonyme (insgesamt)	*75 472*	*11 363*	*10 877*	*97 712*
⌀ Hyperonyme (pro Bedeutung)	*1.054446*	*1.030564*	*1.021794*	*1.047894*
Σ Synonyme (insgesamt)	*65 718*	*9 788*	*9 086*	*84 592*
⌀ Synonyme (pro Bedeutung)	*0.918169*	*0.887719*	*0.853546*	*0.907191*
Σ Antonyme (insgesamt)	*1 504*	*492*	*1 486*	*3 482*
⌀ Antonyme (pro Bedeutung)	*0.021012*	*0.044621*	*0.139596*	*0.037342*

Tabelle 2: Relationen in GermaNet

Relation	Substantive	Verben	Adjektive	Insgesamt
Σ Hyponyme (insgesamt)	*41 726*	*2 018*	*1 597*	*45 341*
⌀ Hyponyme (pro Bedeutung)	*0.447565*	*0.138153*	*0.119322*	*0.374039*
Σ Hyperonyme (insgesamt)	*66 956*	*3 435*	*2 093*	*72 484*
⌀ Hyperonyme (pro Bedeutung)	*0.718189*	*0.235161*	*0.156381*	*0.597954*
Σ Synonyme (insgesamt)	*40 340*	*11 435*	*8 927*	*62 045*
⌀ Synonyme (pro Bedeutung)	*0.432698*	*0.782844*	*0.666990*	*0.289952*
Σ Antonyme (insgesamt)	*21 806*	*4 415*	*10 270*	*35 148*
⌀ Antonyme (pro Bedeutung)	*0.233897*	*0.302252*	*0.767334*	*0.511838*

Tabelle 3: Relationen in Wiktionary (nativ)

Im ersten Vergleich zwischen dem nativen Wiktionary-Wortnetz und GermaNet fällt auf, dass trotz der insgesamt geringeren Anzahl an Synsets in GermaNet mehr

Relation	Substantive	Verben	Adjektive	Insgesamt
Σ Hyponyme (insgesamt)	*222 745*	*14 618*	*6 194*	*243 557*
⌀ Hyponyme (pro Bedeutung)	*2.389224*	*1.000753*	*0.462791*	*2.009215*
Σ Hyperonyme (insgesamt)	*115 822*	*9 325*	*3 969*	*129 116*
⌀ Hyperonyme (pro Bedeutung)	*1.242339*	*0.638393*	*0.296548*	*1.065138*
Σ Synonyme (insgesamt)	*121 616*	*48 644*	*29 710*	*199 970*
⌀ Synonyme (pro Bedeutung)	*1.034487*	*3.330184*	*1.822624*	*1.649645*
Σ Antonyme (insgesamt)	*54 769*	*18 775*	*24 394*	*97 965*
⌀ Antonyme (pro Bedeutung)	*0.587757*	*1.285343*	*2.219815*	*0.808159*

Tabelle 4: Relationen in Wiktionary (symmetrisch)

Relationskanten zwischen den Synsets eingetragen sind als im Wiktionary-Wortnetz. Dementsprechend ist der Unterschied pro Bedeutung (also den einelementigen Synsets in Wiktionary) noch größer. Ausnahme hierbei sind die Antonymrelationen, von denen sich insgesamt über zehn mal so viele im Wiktionary-Wortnetz finden. Dennoch weist GermaNet über alle Relationen insgesamt ungefähr 30% mehr Kanten auf.
Der Unterschied zwischen den Hypo- und Hyperonymrelationen im Wiktionary-Wortnetz ist jedoch ein erstes Indiz dafür, dass Relationen nicht vollständig utilisiert sind. So ist unklar, wie vielen der Hyponymrelationen eine entsprechende Hyperonymrelation entgegen gesetzt ist und bei wie vielen das nicht der Fall ist. Gleiches gilt dabei auch für Synonym- sowie Antonymrelationen, bei denen nicht zwingend eine entgegengesetzte Kante existiert.
Wie häufig die Relationen asymmetrisch eingetragen sind, lassen die Zahlen aus Tabelle 4 vermuten. In dem symmetrischen Wiktionary-Wortnetz sind über drei mal so viele Relationen kodiert wie im (teilweise) asymmetrischen Pendant. Naiv betrachtet kann die Anzahl der Relationen nicht um mehr als das Doppelte ansteigen, da im schlimmsten Fall jede Relation ohne symmetrische Gegenkante eingetragen ist. Jedoch lassen sich die Relationen im Wiktionary-Wortnetz nicht eindeutig zu einer symmetrischen Relation umwandeln (siehe 2.2.3), wodurch häufig aus einer Relationskante mehrere Gegenkanten entstehen. Der Grad der Asymmetrie im nativen Wiktionary-Wortnetz wird weiter in Kapitel 3.2 untersucht.

Trotz der Ungenauigkeiten durch die Asymmetrie im nativen Wiktionary-Wortnetz zeichnet sich ein klarer Unterschied bei der Hypo- und Hyperonymverkettung für Verben und Adjektive ab. Etwa 80% weniger Relationen weisen auf eine deutlich schlechtere Vernetzung im Wiktionary-Wortnetz hin.
Dies ist vor allem den künstlichen Konzepten in GermaNet geschuldet. Sie erlauben deutlich strukturiertere Hypo- beziehungsweise Hyperonymrelationen zwischen Verben und Adjektiven, die selten direkte Ober/- Unterbegriffe zueinander sind.
So finden sich in GermaNet 119 künstliche Konzepte unter den Verben und 133 unter den Adjektiven. Dabei sind 1 155 Verb-Synsets und 3 471 Adjektiv-Synsets Unterbegriff eines künstlichen Konzepts. Es zeigt sich wie stark die Strukturierung der Verben und besonders der Adjektive von den künstlichen Konzepten abhängt: Über 10% der Verb-Synsets und mehr als 32% der Adjektiv-Synsets werden in Abhängigkeit eines künstlichen Konzepts in die Hypo- und Hyperonymhierarchie eingeordnet.

3.1.3 Bedeutungen

Beim Vergleich der Anzahl der Terme (siehe Abbildung 5), fällt auf, dass sich in GermaNet rund 25% mehr Terme befinden als im Wiktionary-Wortnetz. Die geringere Anzahl an Termen bei einer höheren Anzahl an Synsets im Wiktionary-Wortnetz (siehe Kapitel 3.1.1) deutet darauf hin, dass jeder Term mehr unterschiedliche Bedeutungen trägt als in GermaNet. Im Folgenden wird vom **Grad der Polysemie** die rede sein, wobei dieser der Anzahl an Bedeutungen für einen Term entspricht.
Um den Grad der Polysemie pro Term zu vergleichen, wird zunächst die Schnittmenge an Termen gebildet, die sowohl in GermaNet als auch im Wiktionary-Wortnetz vorkommen (mehr dazu in Kapitel 3.3). Dann wird für jeden Term geprüft, in welchem der beiden Wortnetze mehr Bedeutungen kodiert sind. Für GermaNet bedeutet das also, dass für ein Lexem die Synsets gezählt werden, in denen es vorkommt. Die Ergebnisse sind in Abbildung 2 dargestellt.

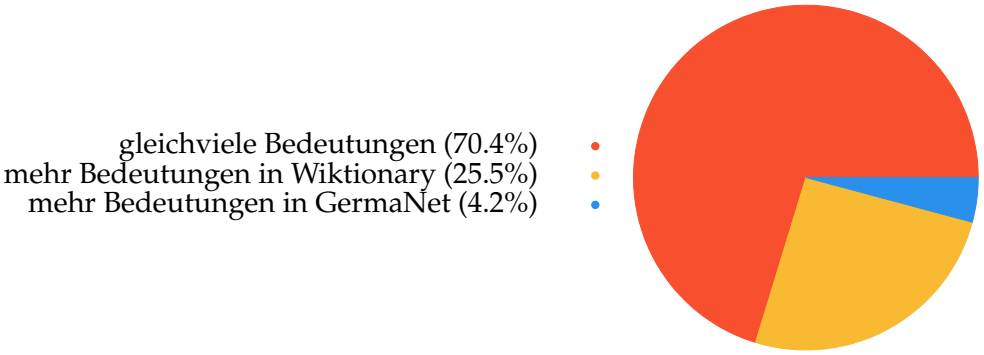


Abbildung 2: Unterschiede im Grad der Polysemie pro Term

Wie deutlich zu erkennen ist, sind für einen großen Teil der Terme mehr Bedeutungen im Wiktionary-Wortnetz eingetragen. Die große Zahl an Termen mit gleichem Grad der Polysemie ist zu über 91% Termen geschuldet, für die nur eine einzige Bedeutung eingetragen ist. Es ist davon auszugehen, dass die meisten dieser Begriffe auch tatsächlich monosem sind und nicht nur unvollständig in Wiktionary eingetragen sind. Das Ergebnis ist besonders bemerkenswert, da beim Parsen bereits feinere Aufteilungen der Bedeutungen zusammengefasst wurden (siehe Kapitel 2.2.1).

3.2 Qualität der Vernetzung

Wie sich in Kapitel 3.1 bereits angedeutet hat, sind die Einträge im Wiktionary inkonsistent. So hat sich herausgestellt, dass ein Großteil der Relationen asymmetrisch kodiert ist, obwohl jede im geparsten Wiktionary-Wortnetz eingetragene Relation symmetrisch sein sollte. Das Ausmaß der Asymmetrie sowie Inkonsistenzen bei der Hypo- und Hyperonymie Strukturierung werden im Folgenden untersucht.

3.2.1 Asymmetrie der Relationen

Abbildung 3 zeigt die prozentual asymmetrisch eingetragenen Relationskanten im Wiktionary-Wortnetz. Es wird deutlich, wie groß die Anzahl der asymmetrisch eingetragenen Relationen tatsächlich ist. Über 50% aller Relationen sind dabei ohne entsprechende symmetrische Gegenkante eingetragen, wobei besonders Oberbegriffsbeziehungen zu über 65% inkonsistent sind.
Da es wesentlich mehr Oberbegriffs- als Unterbegriffsbeziehungen gibt, ließ sich vermuten, dass es weniger asymmetrische Unterbegriffsbeziehungen gibt. Dennoch zeichnet sich bei den Hyperonymen eine besondere Inkonsistenz ab: Deutlich mehr Hyperonymrelationen sind im Gegensatz zu den Hyponymrelationen asymmetrisch. Da Hyperonymrelationen zusätzlich noch zahlreicher sind als Hyponymrelationen, wird hier die Häufigkeit asymmetrischer Oberbegriffsbeziehungen besonders deutlich.

Obwohl es nur 60% mehr Hyperonymrelationen gibt sind mehr als doppelt so viele dieser Relationen im Vergleich zu den Hyponymen asymmetrisch.

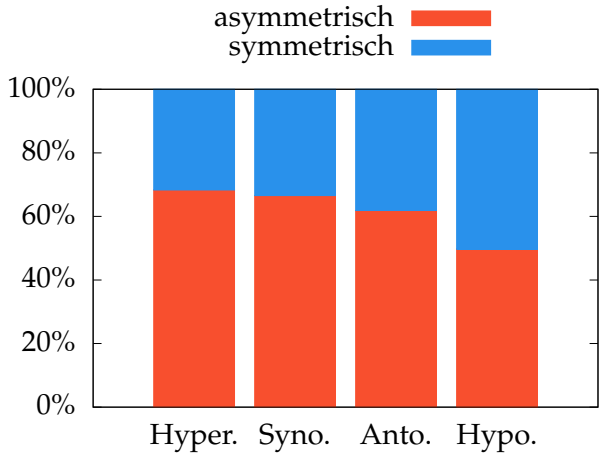


Abbildung 3: Symmetrisch versus asymmetrisch eingetragene Relationen im Wiktionary-Wortnetz

Das Ergebnis zeigt, dass alle Relationstypen im Wiktionary-Wortnetz in den meisten Fällen einseitig eingetragen werden. Insbesondere bei Oberbegriffen sind die Autoren großzügig und vernachlässigen häufig die Gegenrichtung. Es empfiehlt sich also, das Wiktionary symmetrisch in seinen Relationen zu parsen. Wie sich im Folgenden Kapitel zeigen wird, bringt dies aber wieder andere Probleme mit sich.

3.2.2 Inkonsistenz der Hypo- und Hyperonymrelationen

Wie bereits diskutiert, sind die Autoren in Wiktionary nicht immer konsistent beim Eintragen von Relationen. Die Ober- und Unterbegriffsbeziehungen stellen insofern eine besondere Herausforderung dar, da sie die eingetragenen Terme in einen großen, im Wiktionary schwer erkenntlichen, Kontext einordnen müssen. Im Gegensatz zu

Synonymen und Antonymen ist für diese Relationen ein größeres Bild der Struktur zwischen den Termen nötig.
Es stellt sich also die Frage, wie viele Fehler bei der Strukturierung der Hypo- und Hyperonyme gemacht wurden und wie sich diese auf das Wiktionary-Wortnetz auswirken. Hierzu wird das Wortnetz nach einfachen Inkonsistenzen in der Hypo- und Hyperonymhierarchie untersucht. In Abbildung 4 findet sich ein exemplarisches Beispiel aus dem Wiktionary.

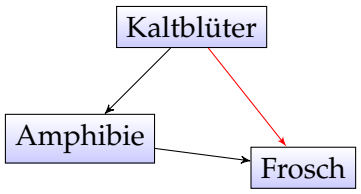


Abbildung 4: Inkonsistente Hyponymrelation (rot)

Gesucht wird nach einer Unterbegriffsbeziehung, die von einem Unterbegriff B (Amphibie) eines Terms A (Kaltblüter) zu einem anderen Unterbegriff C (Frosch) des gleichen Terms A führt. Gleiches wird aufgrund der Asymmetrie ebenso für Hyperonyme untersucht. Dies ist insofern inkonsistent, da (exemplarisch) „Frosch“ kein Unterbegriff von „Kaltblüter“ sein sollte, wenn er schon Unterbegriff von „Amphibie“ ist, welches dann die Hyponymie zu „Kaltblüter“ impliziert.
Solche **Inkonsistenzen** treten vermutlich auf, wenn etwa (am Beispiel von Abbildung 4) der Term Amphibie nach den Termen Kaltblüter und Frosch eingetragen wurde. Amphibie wurde als Hyponym zu Kaltblüter eingetragen und Frosch als Hyponym zu Amphibie, aber das Löschen von Frosch als Hyponym zu Kaltblüter wurde vernachlässigt.

Die Untersuchung zeigte, dass etwa 3,6% der im Wiktionary-Wortnetz vorkommenden Terme in einer inkonsistenten Hyponymhierarchie eingetragen sind und sogar über 12,5% in einer inkonsistenten Hyperonymhierarchie. Wieder zeichnet sich eine größere Inkonsistenz der Hyperonymrelationen ab.
Wie in Kapitel 3.2.1 bereits festgestellt, empfiehlt sich, aufgrund der häufigen Asymmetrie der Relationen, eine symmetrische Betrachtung des Wiktionarys. Diese lässt sich unter Kompromissen erzeugen (siehe Kapitel 2.2.3). Während hierdurch zwar die Asymmetrie behoben wird verstärken sich abseits des Präzisionsverlustes der Relationen aber noch andere Probleme.

Tabelle 5 zeigt die Anzahl der hierarchisch inkonsistent eingetragenen Terme im nativen und im symmetrischen Wiktionary-Wortnetz. Wie deutlich zu erkennen ist, vervielfältigt sich die hierarchische Inkonsistenz abhängig von sowohl den Hypo- als auch den Hyperonymrelationen. Dies führt letztlich dazu, dass über 23% der hierarchischen Relationen im symmetrischen Wiktionary-Wortnetz inkonsistent sind, was vor allem gegenüber den nur 3,6% der inkonsistenten Hyponymrelationen eine Verschlechterung ist.

	nativ	sym.
Hyponym	*3 164*	*19 883*
Hyperonym	*10 836*	*19 883*

Tabelle 5: Terme mit inkonsistenten Hypo-/ Hyperonymrelationen im nativen und symmetrischen Wiktionary-Wortnetz

Die autorenbedingten Fehler und Inkonsistenzen beim Eintragen der Relationen fallen letztlich gegenüber dem wohlstrukturierten und in sich vollständigen GermaNet ins Gewicht. Besonders die Asymmetrie streicht viele theoretisch schon vorhandene Relations-

informationen aus dem Wiktionary-Wortnetz, wobei das programmatische Hinzufügen fehlender Relationen unter anderem zu Problemen mit der Hierarchie führt.

3.3 Abdeckung

In diesem Kapitel wird sich den Termen gewidmet, die in den Wortnetzen eingetragen sind. Zunächst wird Schnittmenge und Differenz der Terme aus dem Wiktionary und GermaNet bestimmt, um die Analogie der beiden Wortnetze zu untersuchen. Dann wird die Abdeckung der Terme aus verschiedenen repräsentativen Wortmengen untersucht, so dass die Abdeckungsbereiche der beiden Wortnetze genauer skizziert werden können. Zudem werden die in Wiktionary teilweise eingetragenen Domänen geparst und die kategorisierten Lexeme weiter betrachtet.

3.3.1 Differenz und Schnittmenge

Die erste und einfachste Frage, die sich beim Vergleich der in den Wortnetzen abgedeckten Terme stellt ist, inwiefern die beiden Wortnetze sich grundsätzlich ähneln beziehungsweise worin sie sich unterscheiden. Dazu lässt sich der in Abbildung 5 veranschaulichte Schnitt der beiden Term-Mengen betrachten.

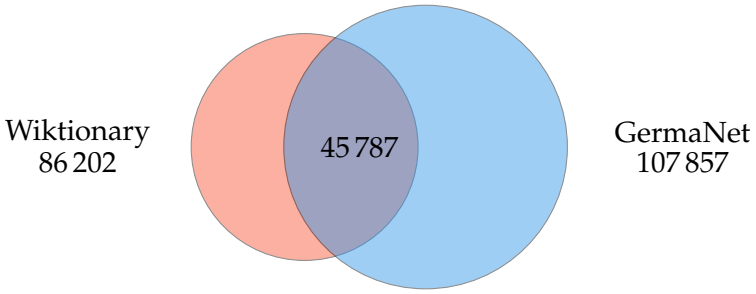


Abbildung 5: Differenz und Schnitt der im Wiktionary-Wortnetz (rot) und GermaNet (blau) eingetragenen Terme.

Gerade einmal etwa 53% der im Wiktionary-Wortnetz eingetragenen Terme sind auch in GermaNet eingetragen, welches sogar nur etwa 42% seiner Terme mit dem Wiktionary-Wortnetz teilt. Es wird deutlich, dass beide Wortnetze zu großen Teilen disjunkt sind, wobei GermaNet aufgrund seines Umfangs mehr Terme aus dem Wiktionary-Wortnetz abdeckt als das Wiktionary-Wortnetz Terme aus GermaNet.
Besonders bei den Verben umfasst GermaNet eine Mehrheit der in Wiktionary eingetragenen Terme. Über 75% der Verben aus Wiktionary sind auch in GermaNet eingetragen.

Insgesamt sind 102 520 der 194 094 Terme, die beide Netze zusammen aufbringen, nur im Wiktionary-Wortnetz oder nur in GermaNet zu finden. Es stellt sich also die Frage, welche Terme in den Wortnetzen einzigartig sind. Dies wird weiter in den folgenden Kapiteln untersucht.

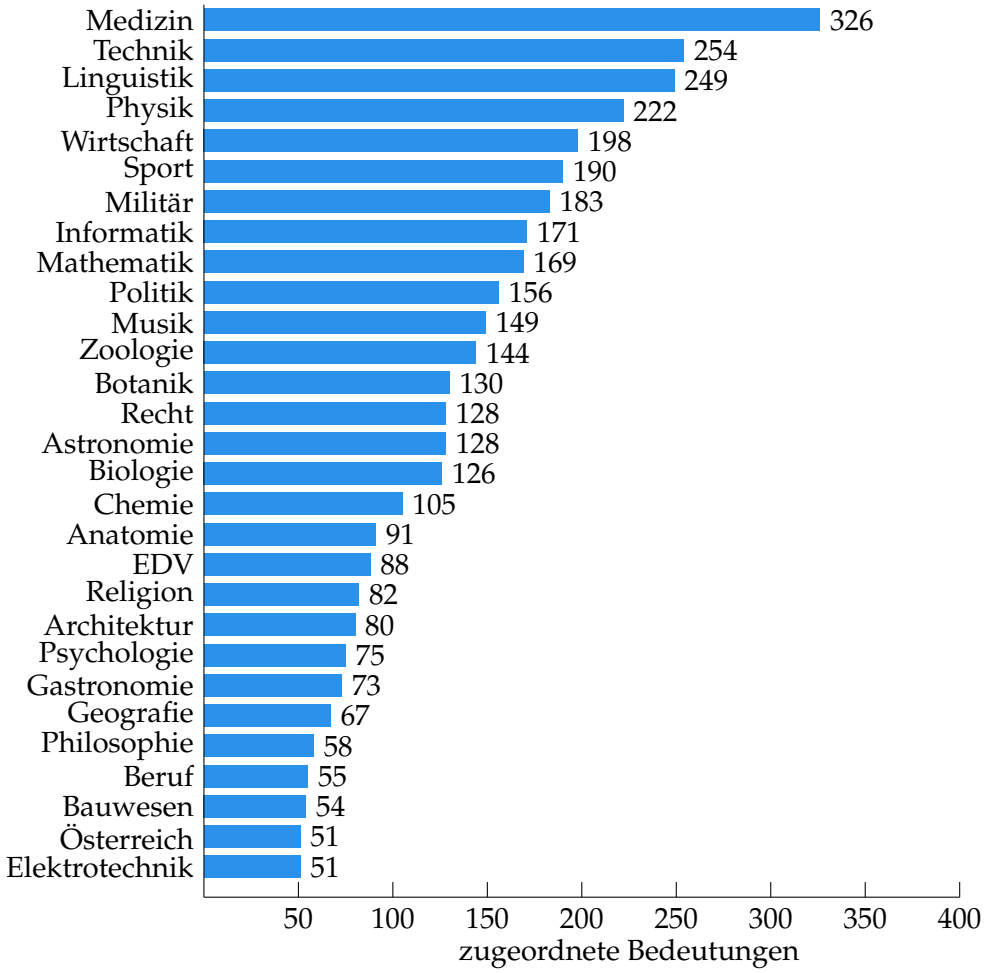


Abbildung 6: Domänen mit mehr als 50 im Wiktionary-Wortnetz zugeordneten Bedeutungen

3.3.2 Wiktionary Domänen

Wiktionary bietet den Autoren die Möglichkeit, Bedeutungen eines Terms mit einer Domäne zu kennzeichnen (siehe Zeile 15 und 17 in Quellcode 1). **Domänen** sind hierbei Fach- beziehungsweise Themenbereiche, in welchen der jeweilige Term gebraucht wird. Diese Kennzeichnung wird sich zu Nutze gemacht, indem sie zusätzlich geparst und dann den entsprechenden Lexemen zugeordnet wird.
Da grundsätzlich keine Regelung zu Domänen besteht, steht es den Autoren frei welche Informationen sie in die Domänenkennzeichnung eintragen. So ist es etwa möglich, dass eine Bedeutung mehreren Domänen zugeordnet ist. Auch müssen die Ergebnisse nach sinnvollen Domänen gefiltert werden. Schließlich ergaben sich 29 zum Vergleich geeignete Domänen, deren Quantität in Abbildung 6 dargestellt ist.

Auch wenn die Anzahl der mit einer Domäne versehenen Bedeutungen verhältnismäßig gering ist, geben die Ergebnisse einen groben Überblick darüber, welche Themenberei-

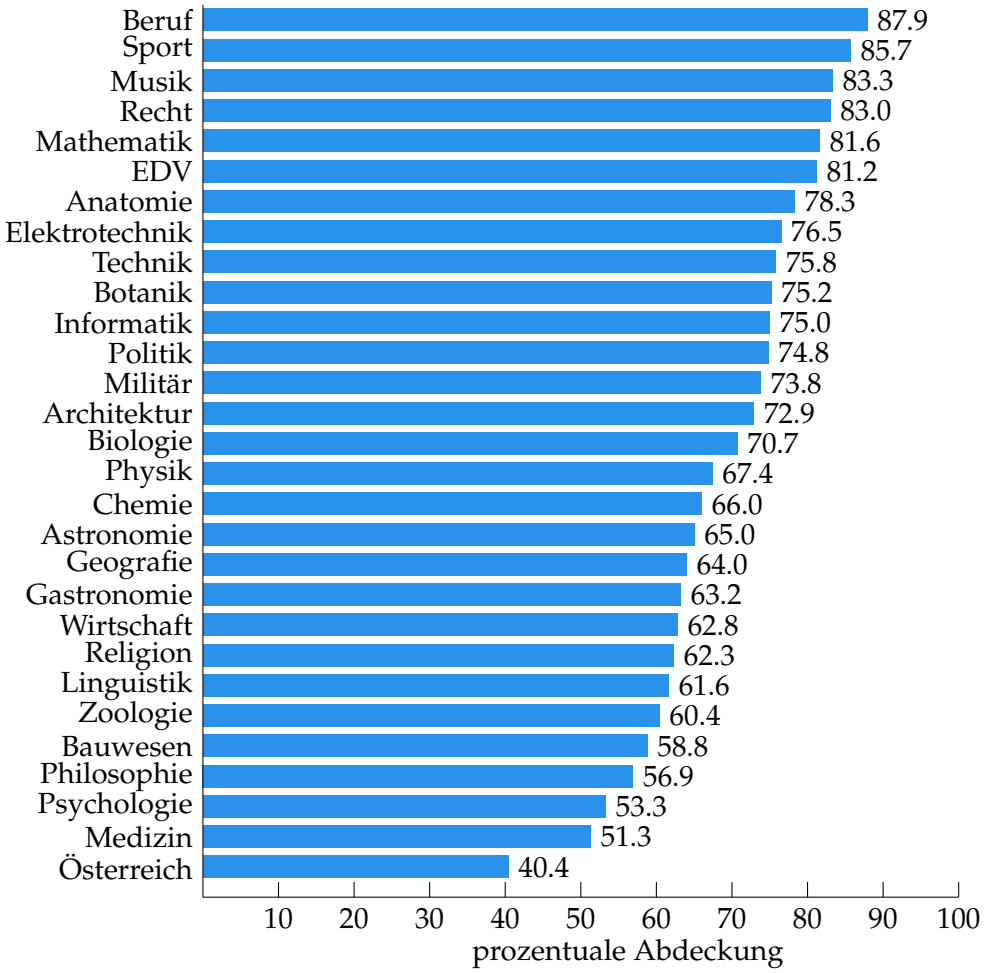


Abbildung 7: Prozentuale Abdeckung der mit Domänen gekennzeichneten Terme aus dem Wiktionary in GermaNet

che im Wiktionary besonders vertreten sind. Heraus sticht vor allem der medizinische Bereich mit mehr als 300 verzeichneten Bedeutungen. Auch technische und linguistische Lexeme sind häufig vertreten.
Es ist zu beachten, dass die Filterung untergruppierte Domänen nicht zu ihrer übergeordneten Domäne zuordnet. So findet sich beispielsweise Elektrotechnik unter den qualifizierten Domänen, obwohl Lexeme, die mit Elektrotechnik klassifiziert sind, ebenfalls zu der Domäne Technik gehören sollten. Dies lässt vermuten, dass die Zahl der zugeordneten Lexeme teilweise sogar noch höher ist, viele jedoch spezifischeren, untergeordneten Domänen zugeordnet sind.

Ein direkter Vergleich der mit Domänen gekennzeichneten Terme aus dem Wiktionary-Wortnetz mit GermaNet liefert Aufschluss darüber, wie die einzelnen Themenbereiche in GermaNet vertreten sind. Die Ergebnisse sind in Abbildung 7 dargestellt. Für die meisten Domänen zeigt sich ein vorhersehbares Ergebnis von 60% bis 80% Vorkommen in GermaNet, welches aufgrund der höheren Abdeckung von Wiktionary-Termen in GermaNet

(siehe Kapitel 3.3.1) zu erwarten war.
Besonders Berufsbezeichungen aus der Domäne Beruf sind zu fast 90% in GermaNet vertreten. Lediglich österreichische Terme wie „Hausbesorger“ oder „Wissenschafter“ und alte, spezielle Terme wie „Knopfmacher“ finden sich dort nicht.
Wie auch schon die große Anzahl an mit Medizin gekennzeichneten Bedeutungen vermuten lies, sind in Wiktionary gegenüber GermaNet besonders viele medizinische Terme verzeichnet. Die auf Wiktionary beschränkten Terme sind zumeist fachspezifisch (etwa „Polyurie“, „Abusus“ oder „Laryngitis“), während die mit GermaNet geteilten Terme eher universeller Natur sind (beispielsweise „Schmerz“, „Atemweg“ oder „Rückbildung“).

Besonders gering ist die Abdeckung österreichischer Begriffe. Lediglich etwa 40% der im Wiktionary eingetragenen Terme befinden sich auch in GermaNet. Um die Abdeckung der österreichischen Terme in GermaNet genauer zu untersuchen wurde die Filterung nach österreichischen Domänen weiter verfeinert (Terme aus Domänen wie „in Österreich“, „westösterreichisch“ oder „österr.“ wurden zu Österreich hinzugezählt).
So findet sich eine deutlich höhere Anzahl an österreichischen Termen (165 gegenüber vorher 47), die prozentuale Abdeckung dieser Terme in GermaNet bleibt jedoch mit 40% etwa gleich gering. Ähnliche Ergebnisse für schweizerdeutsche Terme deuten auf eine generell bessere Abdeckung der alemannischen Dialekte im Wiktionary hin.

3.3.3 Gut1 Grundwortschatz 500

Nachdem nun die Analogie zwischen GermaNet und dem Wiktionary-Wortnetz geprüft wurde, wird die Abdeckung verschiedener repräsentativer Wortmengen bestimmt um den Umfang und gegebenenfalls Spezialisierungsgebiete zu ermitteln.
Den Anfang macht der Gut1 Grundwortschatz 500[12]. Die Abdeckung der in dieser Zusammenstellung aus den 500 grundlegensten Wörtern der deutschen Sprache gibt Aufschluss über mögliche Mängel bei der Abdeckung elementarer Terme.

Es stellt sich jedoch heraus, dass in keinem der beiden Wortnetze Lücken bestehen. Zwar decken Wiktionary und GermaNet nur 68,8% beziehungsweise 72,2% des Gut1 Wortschatzes ab, jedoch umfasst dieser auch Pronomen, Adverbien, Konjunktionen und Präpositionen vor, welche beide Wortnetze nicht abdecken.
Der Unterschied zwischen GermaNet und Wiktionary kommt hauptsächlich daher, dass Numerale in GermaNet als Substantive („Eins“ wie die Klausurnote oder „Nächste“ wie eine nahestehende Person) beziehungsweise als Adjektive („alle“, „kein“, „neun“) eingetragen sind.
Die einzigen Terme, die sich zwar in Wiktionary, nicht jedoch in GermaNet finden, sind „hinter“, „zu“ und „gar“, welche alle drei als Adjektive in Wiktionary eingetragen sind.

Es ist jedoch zu beachten, dass ausnahmslos alle Terme einen Eintrag im Wiktionary besitzen, nur nicht in das Wortnetz geparst wurden, da sie nicht von gewünschter Wortart sind. Durch das Erweitern des Parsers würden also auch diese Terme im Wiktionary-Wortnetz übernommen werden.
Abhängig davon, wie viel Nutzen sich aus Termen abseits der Substantive, Verben und

[12] https://www.gut1.de/grundwortschatz/grundwortschatz_500.html

Adjektive ziehen lässt, findet sich mit leicht erhöhtem Parsing-Aufwand im Wiktionary-Wortnetz ein umfangreicherer Kandidat.

3.3.4 Projekt deutscher Wortschatz

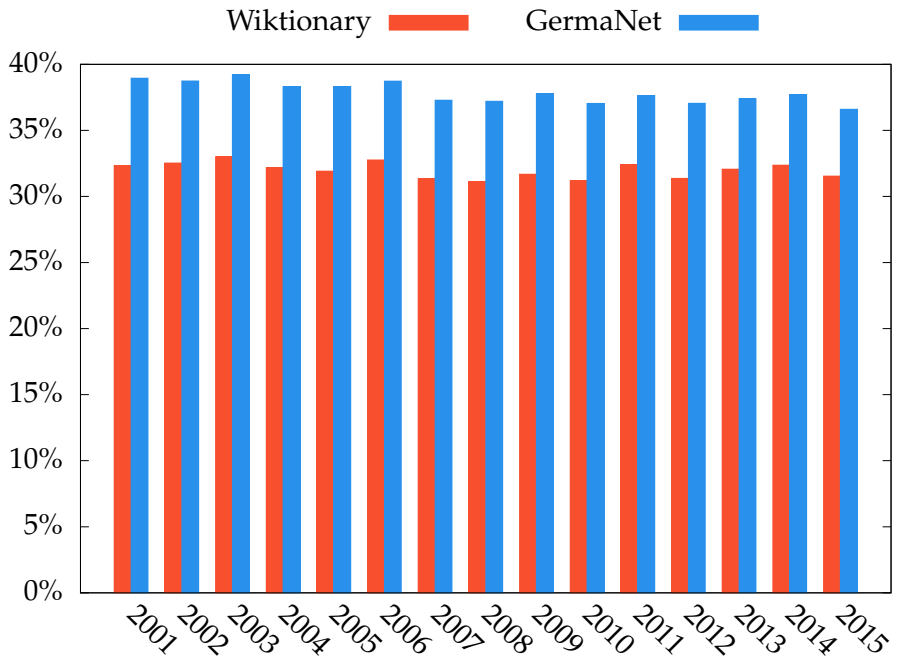


Abbildung 8: Abdeckung der PdW Terme in Wiktionary und GermaNet pro Jahr

Wie in Kapitel 3.3.3 deutlich geworden, ist der Grundwortschatz von beiden Wortnetzen abgedeckt. Es gilt also nun die Wortnetze auf ihren weitreichenderen Umfang zu testen. Das Projekt deutscher Wortschatz (kurz: *PdW*) der Universität Leipzig [GEQ12] erstellt seit 1995 für jedes Jahr einen Korpus aus zufällig ausgewählten Sätzen aus Nachrichtentexten und allgemeinen Webcrawlings. Damit enthält es viele Terme wie sie für die spätere Anwendung des Wiktionary-Wortnetzes möglicherweise auch gebraucht werden. Zum Extrahieren dieser Termmenge wird der Tokenizer aus *spaCy*[13] genutzt, der über die Sätze der deutschen - jeweils 10 000 Sätze umfassenden - Datensätze aus dem PdW läuft und die Wortart der Wörter identifiziert. Daraus ergibt sich, gefiltert nach Substantiven, Verben und Adjektiven, eine Menge aus etwa 28 000 Termen (pro Jahr). Die Abdeckung dieser Wörter wird im Wiktionary-Wortnetz und GermaNet untersucht.

Wie Abbildung 8 deutlich zeigt, finden sich in GermaNet über alle Datensätze mehr der in den Nachrichtentexten und Webcrawlings auftauchenden Terme. Dies war jedoch aufgrund des größeren Umfangs GermaNets zu erwarten.
Dennoch scheint das Wiktionary-Wortnetz, trotz seiner geringeren Abdeckung der PdW Terme, insgesamt besser auf die Terme aus dem Datensatz zugeschnitten. GermaNet besteht aus etwa 25% mehr Termen, deckt aber, selbst im Jahr der größten Differenz, nur

[13] `https://spacy.io/`

20% mehr Terme ab als Wiktionary. Im aktuellsten Datensatz aus 2015 sind es sogar nur noch etwa 16% mehr Terme die sich zwar in GermaNet, jedoch nicht im Wiktionary-Wortnetz nachschlagen lassen.

In Abbildung 9 wird die Differenz der prozentualen Abdeckungsergebnisse $G_{\text{PdW-Abd.}} - W_{\text{PdW-Abd.}}$ dargestellt, wobei $G_{\text{PdW-Abd.}}$ und $W_{\text{PdW-Abd.}}$ die prozentuale Abdeckung der PdW Terme in GermaNet beziehungsweise Wiktionary sind. Die resultierende Kurve zeigt, dass sich die Abdeckung in Wiktionary im Verhältnis zu GermaNet über die Jahre verbessert.
Grund hierfür ist möglicherweise Wiktionarys Aktualität. Während das kollaborative Wörterbuch nach seinem Start im Mai 2004 vermutlich zunächst mit den gundlegenden deutschen Termen gefüllt wird, beginnen die Nutzer ab 2009 auch weniger elementares Vokabular zu dokumentieren. Die erhöhte Anzahl an Einträgen zu aktuellen Begriffen, wie etwa Neologismen, verschafft dem Wiktionary-Wortnetz einen Vorteil gegenüber GermaNet, welches nur langsam in jährlichen Zyklen aktualisiert wird.

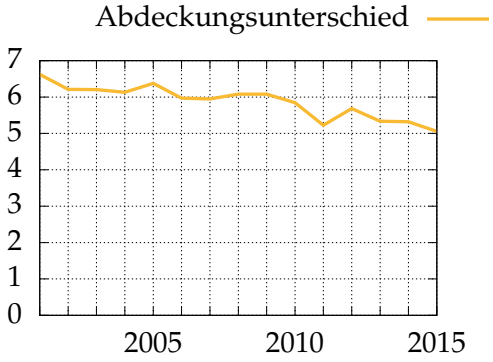


Abbildung 9: Differenz der prozentualen Abdeckung der PdW Terme im Wiktionary-Wortnetz und GermaNet pro Jahr

Die Abdeckung der PdW Terme zeigt einmal mehr, dass GermaNet grundsätzlich eine umfangreichere Datenbank bietet. Wiktionary hingegen ist, ähnlich wie in Kapitel 3.3.2, ein guter Kandidat für speziellere Terme, wie in diesem Fall aktuelle, zeitnahe Terme.

3.3.5 Neologismen

Einer der Vorteile von kollaborativen Wissensbasen ist das schnelle Wachstum sowie die zügige Adaption an neue Begriffe. Besonders in der Zeit des Internets verbreiten sich Neologismen rasch und werden schnell in den allgemeinen Sprachgebrauch übernommen. Damit ist die Aktualität eines Wortnetzes ein maßgebliches Kriterium, wenn es darum geht, *natural language processing* auf gegenwartsnahen Daten anzuwenden.
Für die Untersuchung werden zwei Quellen für Neologismen herangezogen. Zunächst wird dabei die Website *kunst-worte.de*[14] behandelt, die es registrierten Nutzern ermöglicht, Neologismen der deutschen Sprache zusammen mit einer deskriptiven Erläuterung zu verzeichnen. Aus den eingetragenen Neologismen ergab sich eine Menge von 150 Termen, deren Abdeckung nun im Wiktionary-Wortnetz und GermaNet untersucht wird.

Die Untersuchung zeigt, dass Wiktionary tatsächlich eine deutlich bessere Abdeckung der Neologismen von *kunst-worte.de* liefert. Lediglich drei der geprüften Neologismen finden sich gegenüber dem Wiktionary-Wortnetz ausschließlich in GermaNet: „Workaround“, „Gadget“ und „Augenkrebs“, wobei für „Gadget“ ein mit englischer Herkunft

[14] `https://www.kunst-worte.de/`

gekennzeichneter Eintrag im deutschen Wiktionary besteht (der Parser ist auf deutsche Einträge beschränkt). Besonders hochaktuelle Internettermini wie „Shitstorm“ oder „Emoji“ sind bereits im Wiktionary-Wortnetz vertreten. Aber auch aktuelle kontextuelle Neologismen wie etwa „Nafri“ sind bereits verzeichnet.
Insgesamt bleibt die prozentuale Abdeckung der 150 Terme jedoch gering: 34,44% der Neologismen deckt das Wiktionary ab, wo hingegen 11,92% von GermaNet abgedeckt werden. Die Abdeckung im Wiktionary-Wortnetz bleibt unter Anderem auch deshalb so gering, da viele der deutschen Neologismen aus dem Englischen kommen, eine analoge Übersetzung sind oder sich aus dem englischen Neologismus ableiten (beispielsweise das Verb „fabbing“, welches die dreidimensionale Produktion von Objekten mittels 3D-Druckern beschreibt). Für manche dieser Neologismen existiert nur ein Eintrag für den englischen Term, wodurch dieser beim Parsen ebenfalls nicht in das Wortnetz aufgenommen wurde (wie etwa „scam“).
Bei anderen Neologismen, wie dem auf *kunst-worte.de* als Substantiv aufgefassten Term „Bufti“, handelt es sich um eine andere Schreibweise beziehungsweise Fehlinterpretation der Abkürzung „Bufdi“, für die sich wieder im Wiktionary ein Eintrag findet, der beim Parsen nicht übernommen wurde (da Abkürzung und damit nicht geparste Wortart). In GermaNet fehlt der Term „Bufdi“ jedoch gänzlich.

Die starken Unterschiede bei der Abdeckung der Neologismen zeichnen den Vorteil der zügigen Aktualisierung Wiktionarys ab. Auch hier lässt sich durch Anpassung des Parsings der Term-Umfang des Wiktionary-Wortnetzes noch weiter solidieren, um etwa auch Abkürzungen oder Anglizismen besser abzudecken.

Eine weitere, wesentlich umfangreichere Quelle für deutsche Neologismen ist, mit über 56 000 Einträgen, die Website *wortwarte.de*[15], deren Begriffe für diese Arbeit über einen Webcrawler extrahiert wurden. Das Projekt dokumentiert meist täglich neu aufgekommene Begriffe mit einer Quelle, in der diese auftauchen. Die dadurch entstehende zeitliche Zuordnung der Neologismen ermöglicht zusätzlich eine Untersuchung der Terme in Abhängigkeit ihrer Aktualität.

Die Ergebnisse dieser Untersuchung sind in Abbildung 10 dargestellt. In erster Linie fällt die deutlich bessere, 2001 sogar fast vier mal so hohe Abdeckung der Neologismen in GermaNet für den Zeitraum von 2000 bis 2003 auf. Hier ist allerdings davon auszugehen, dass Begriffe die zu dieser Zeit neu waren, nicht mehr als Neologismus zählen. Viele der Terme sind schon lange im deutschen Sprachgebrauch üblich. So finden sich in diesem Zeitraum etwa Terme wie „Chatraum“, „mailen“ oder „meistgeklickt“, welche alle im heute üblichen Internet bezogenen Sprachgebrauch vorkommen.
Beginnend mit 2004 verändert sich das Bild jedoch rapide. Die Abdeckung der Neologismen in GermaNet ist dramatisch schlechter, während sie zumindest von 2003 auf 2004 in Wiktionary noch ein wenig steigt. In allen darauf folgenden Jahren ist die Abdeckung im Wiktionary-Wortnetz besser als die des GermaNets, jedoch insgesamt durchgehend sehr gering.

Die insgesamt schlechte Abdeckung der Neologismen ist, wie schon für die Terme der Website *kunst-worte.de* diskutiert, zum Teil auch unterschiedlichen Schreibweisen oder Termen abseits der in den Wortnetzen vertretenen Wortarten geschuldet. Dazu kommt

[15] `http://www.wortwarte.de/`

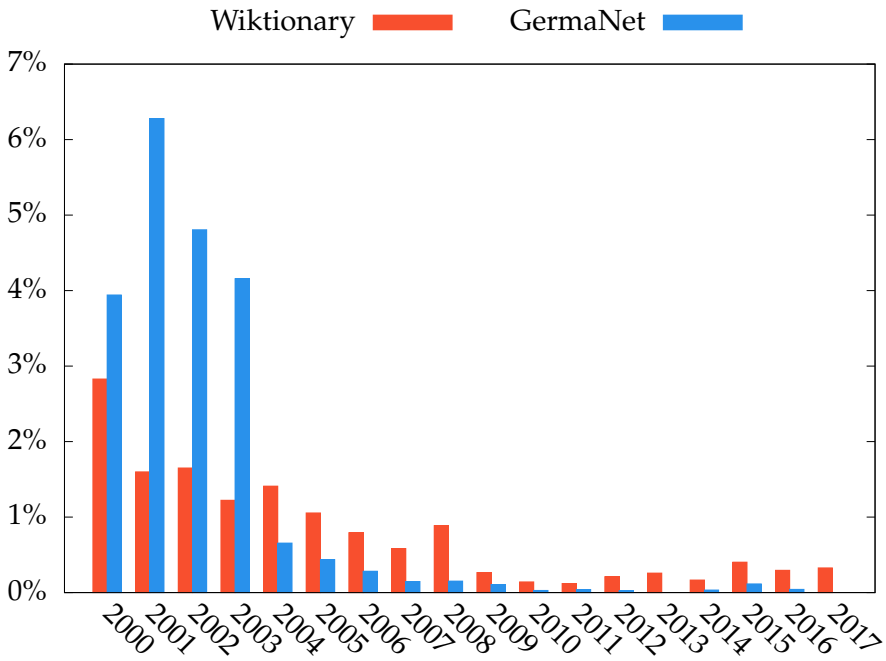


Abbildung 10: Abdeckung der Wortwarte Terme in Wiktionary und GermaNet pro Jahr

für die Terme der *Wortwarte*, dass viele der eingetragenen Begriffe besonders speziell sind oder nur sehr kurzlebig waren, was eine Aufnahme in eines der beiden Wortnetze zusätzlich unwahrscheinlicher macht.
Ein Beispiel hierfür wäre etwa der „Brokkolistreit“: In einem Artikel veröffentlicht auf Spiegel Online[16] berichtet die Autorin über einen gerichtlichen Streit um ein Brokkolipatent, den sie schließlich mit Brokkolistreit betitelt. Der Begriff wurde in der *Wortwarte* eingetragen, jedoch danach nie wieder benutzt.
Anders ist es mit den Begriffen, die Wiktionary tatsächlich abdeckt. Wieder taucht beispielsweise 2015 der Term „Smombie“ auf. Auch andere aktuellere Begriffe wie „Tabletcomputer“ sind unter den Begriffen, die Wiktionary nach 2003 gegenüber GermaNet abdeckt.

Es zeigt sich also auch hier ein eindeutiger Aktualitätsvorteil für Wiktionary. Zwar sind die älteren Begriffe in GermaNet besser abgedeckt, die für diese Untersuchung jedoch relevanten Neologismen sind in Wiktionary eindeutig häufiger vertreten.

3.4 Vergleich über *word embeddings*

Bisher hat sich der Vergleich der beiden Wortnetze überwiegend der Quantität und den Eigenschaften der eingetragenen Terme gewidmet. Um auch die Qualität der Relationen zu Untersuchen wird eine sehr aktuelle Repräsentation von syntaktischen und seman-

[16] `http://www.spiegel.de/wirtschaft/patente-auf-nahrungsmittel-die-brokkoli-revolte-a-707385.html`

tischen Zusammenhängen zwischen Termen zum Vergleich herangezogen: *word embeddings*.

Die Idee hinter *word embeddings* ist das Abbilden von Wörtern oder Phrasen in einen Vektorraum. Somit werden für Terme Vektorrepräsentationen erzeugt, deren Ähnlichkeit anhand der mathematischen Vektorähnlichkeit bestimmt werden kann.
Word embeddings sind besonders seit der Veröffentlichung von Word2vec [MCCD13], einem Toolkit welches den Zugang zum Trainieren von eigenen *embeddings* sowie den Gebrauch von *pre-trained embeddings* deutlich leichter gemacht hat. Die üblichsten Methoden, solche Abbildungen von Wörtern in Vektorräume zu generieren, sind neuronale Netzwerke [MSC+13], ebenfalls angewandt werden aber auch probabilistische Modelle [GCPT07] oder Dimensionsreduktion auf der Wort-Grauwertmatrix [LC13].

Auch in dieser Arbeit wird von einem mit Word2vec trainierten Modell Gebrauch geamcht. Das Modell wurde mit den Inhalten des deutschen Wikipedias[17] trainiert und bildet auf einen 300-dimensionalen Vektorraum ab. In diesem Vektorraum wird die Ähnlichkeit zweier Terme beziehungsweise ihrer Vektoren über die Kosinus-Ähnlichkeit

$$cos(\theta) = \frac{A \cdot B}{\parallel A \parallel_2 \parallel B \parallel_2} = \frac{\sum_{i=1}^{n} a_i \cdot b_i}{\sqrt{\sum_{i=1}^{n}(a_i)^2} \cdot \sqrt{\sum_{i=1}^{n}(b_i)^2}}$$

berechnet, wobei A und B die zu vergleichenden Vektoren und a_i beziehungsweise b_i deren Komponenten sind.

3.4.1 Vergleich innerhalb der Schnittmenge

Im Folgenden werden die über ein *word embedding* bestimmten Zusammenhänge zwischen Termen mit den in den Wortnetzen kodierten Zusammenhängen verglichen. Hierfür werden die in 3.3.1 bestimmte Schnittmenge der beiden Wortnetze betrachtet. Diese Menge an Termen wird zusätzlich dadurch eingeschränkt, dass jeder vorkommende Term auch im *word embedding* enthalten sein muss (eine Vektorrepräsentation besitzt). Bildlich gesprochen wird also zunächst die Schnittmenge der Terme aus GermaNet, Wiktionary und den im *word embedding* trainierten Termen gebildet. Diese Schnittmenge wird fortan mit S bezeichnet.
Anschließend werden für jeden in dieser Schnittmenge vorkommenden Term kosinusähnliche Terme im *word embedding* bestimmt. Hierbei müssen zwei Einschränkungen vorgenommen werden:

1. Die Suche ist beschränkt auf die maximal $N = 10$ ähnlichsten Vektoren
2. Die Kosinus-Ähnlichkeit ξ zweier Vektoren muss mindestens ≥ 0.6 sein.

N ist dabei möglichst klein gewählt. Dies hat den Hintergrund, dass im *word embedding* auch viele fälschlich trainierte Terme wie „tausüg“ oder „xhosa“ vorkommen. Außerdem finden sich unter den Gleichartigen häufig flexionen (wie „Hauses“ unter den Gleichartigen zu „Haus“), die für den Abdeckungsvergleich ebenfalls ungeeignet sind. Diese zusammenhanglosen und unzweckmäßigen Ausdrücke werden folgend **falsche Terme**

[17] `https://de.wikipedia.org/`

	Wiktionary	GermaNet
Σ Anzahl an AG	*34 017*	*50 207*
$\%$ Anzahl an AG	*28.75%*	*42.44%*
$\varnothing$ Anzahl an AG pro Term	*2.89*	*4.09*
$\varnothing\%$ Anzahl an AG pro Term	*34.56%*	*48.36%*

Tabelle 6: Abdeckung der Gleichartigen zu Termen aus S im Wiktionary-Wortnetz und GermaNet. AG steht hier für *abgedeckte Gleichartige*, also alle die Gleichartigen, für die in dem jeweiligen Wortnetz ein Eintrag besteht. Die Werte für „$\%$ Anzahl an AG“ sind der prozentuale Anteil an AG von allen 118 311 gefundenen Gleichartigen. Die Werte für „$\varnothing\%$ Anzahl an AG pro Term“ sind der durchschnittliche prozentuale Anteil an AG von all den Gleichartigen, die pro Term gefunden wurden.

genannt. Versuche mit verschiedenen N ergaben für $N = 10$ einen guten Mittelwert aus möglichst vielen ähnlichen Termen und dabei möglichst wenigen falschen Termen.
Die für ξ gewählte Grenze für die minimale Ähnlichkeit dient der Normierung. Hier wurde ebenfalls mit verschiedenen Werten experimentiert. Eine Grenze bei 0.6 erwies sich dabei als gute Balance zwischen Qualität und Quantität, wobei sich die Quantität auf die Anzahl der ähnlichen Vektoren oberhalb der Grenze und die Qualität auf die Menge an falschen Termen oberhalb der Grenze bezieht.

Um die Beschreibung zu erleichtern wird ein Terminus für die Terme eingeführt, die unter den oben genannten Kriterien in die Menge der zu einem Term ähnlichen Wörter fallen: **Gleichartige**. Die Gleichartigen zum Term „Haus“ wären also alle Wörter, die unter den 10 ähnlichsten Wörtern zu „Haus“ im *word embedding* gehören, wobei jedes dieser Wörter eine Ähnlichkeit von mindestens 0.6 hat.

Zuerst wird die Abdeckung der Gleichartigen pro Wortnetz ermittelt. Geprüft wird also, wie viele der aus dem *word embedding* hervorgegangenen Gleichartigen überhaupt im jeweiligen Wortnetz abgedeckt sind.
In S sind 44 060 Terme enthalten zu denen 118 311 Gleichartige bestimmt wurden. Dies entspricht 7.18 Gleichartigen pro Term. Tabelle 6 zeigt die Abdeckung dieser Gleichartigen in Wiktionary und GermaNet.

Die Abdeckung der Gleichartigen ist in diesem Fall überproportional zum Größenunterschied der beiden Netze besser. Fast 50% mehr Gleichartige sind in GermaNet vertreten. Da die Menge der Gleichartigen jedoch nicht nach falschen Termen gefiltert ist, ist unklar, wie groß deren Anteil ist und wie weit sie die Werte beeinflussen.
Daher werden die Mengen T_W und T_G an abgedeckten Gleichartigen in Wiktionary und GermaNet genutzt, um einen fundierteren Vergleich der beiden Wortnetze auf Relationsebene machen zu können.

Die Mengen T_W und T_G enthalten nun ausschließlich Terme, die im jeweiligen Wortnetz vertreten sind. Untersucht wird, inwiefern der im *word embedding* über Vektorähnlichkeit kodierte syntaktische und semantische Zusammenhang mit dem über Synsets und Relationen kodierten lexikalischen Zusammenhang korreliert. Dabei wird davon ausgegangen, dass ein Großteil der abgedeckten Gleichartigen in einer lexikalischen Relation

	Wiktionary (N)	Wiktionary (S)	GermaNet
Σ Anzahl an RG	*11 821*	*15 591*	*16 673*
$\%$ Anzahl an RG	*34.75%*	*45.83%*	*33.21%*
$\varnothing$ Anzahl an RG pro Term	*0.31*	*0.41*	*0.44*
$\varnothing\%$ Anzahl an RG pro Term	*36.34%*	*38.93%*	*28.87%*

Tabelle 7: Abdeckung der in Relation stehenden abgedeckten Gleichartigen zu Termen aus T_W und T_G im Wiktionary-Wortnetz und GermaNet. RG steht hier für *in Relation stehende abgedeckte Gleichartige*, also alle die Gleichartigen, für die in dem jeweiligen Wortnetz ein Eintrag besteht und die zusätzlich in einer Relation zum jeweiligen Term eingetragen sind. Die Werte für „$\%$ Anzahl an RG“ sind der prozentuale Anteil an RG von allen gefundenen abgedeckten Gleichartigen (siehe Tabelle 6) pro Wortnetz. Die Werte für „$\varnothing\%$ Anzahl an AG pro Term“ sind der durchschnittliche prozentuale Anteil an RG von all den Gleichartigen, die pro Term gefunden wurden.

zu dem jeweils zugehörigen Term steht und somit in den Wortnetzen über eine Relation verbunden sein sollten.

Dementsprechend wird nun untersucht, wie viele der abgedeckten Gleichartigen pro Wortnetz jeweils in einer Relation eingetragen sind. Dies wird im folgenden Beispiel veranschaulicht:
Für den Term „Haus“ wurden mittels des *word embeddings* zehn Gleichartige identifiziert:
{Nachbarhaus, Gartenhaus, Häuschen, Wohnhaus, Landhaus, Stadthaus, Sommerhaus, Anwesen, Familienhaus, Patrizierhaus}
Von diesen Gleichartigen sind sieben im Wiktionary abgedeckt:
{Gartenhaus, Häuschen, Wohnhaus, Landhaus, Stadthaus, Anwesen, Familienhaus}
Und von diesen abgedeckten Gleichartigen stehen im nativen Wiktionary fünf in Relation zu „Haus“ (also sind entweder Synonym, Antonym, Hypo- oder Hyperonym):
{Gartenhaus, Wohnhaus, Landhaus, Stadthaus, Familienhaus}

Die Ergebnisse dieser Untersuchung finden sich in Tabelle 7. In diesem Fall wird zusätzlich zwischen dem symmetrischen und dem nativen Wiktionary-Wortnetz unterschieden, da im symmetrischen Fall mehr Relationskanten vorliegen.
Es fällt auf, dass sowohl im symmetrischen wie auch im nativen Wiktionary-Wortnetz mehr abgedeckte Gleichartige in Relation stehen als in GermaNet. Zwar finden sich mehr in Relation stehende abgedeckte Gleichartige pro Term für die Terme aus GermaNet, jedoch ist dies lediglich der größeren Menge an abgedeckten Gleichartigen geschuldet. Durchschnittlich stehen in Wiktionary prozentual mehr abgedeckte Gleichartige pro Term in Relation.

Das Ergebnis lässt vermuten, dass im Wiktionary-Wortnetz syntaktisch und semantisch verwandte Terme eher in einer Relation zueinander kodiert sind als in GermaNet. Dies deutet auf eine vorteilhaftere Vernetzung hin, in der zueinander in Relation stehende Terme eher verknüpft sind.
Jedoch sind die Ergebnisse, wie bereits diskutiert, aufgrund der falschen Terme ungenau. Daher werden beide Wortnetze zusätzlich mit einer kleinen Menge von ausgesuchten Termen untersucht, deren Gleichartige händisch von falschen Termen bereinigt werden.

3.4.2 Vergleich über ausgewählte Terme

Die Menge der ausgewählten Terme ist eine Teilmenge der in Kapitel 3.4.1 bestimmten Menge S. Diese Menge C (siehe Anhang) besteht aus besonders ambiguen Termen die in beiden Wortnetzen stark vernetzt sind. Davon sind 30 Terme Substantive, 10 Verben und 10 Adjektive.
Da die Gleichartigen händisch auf falsche Terme untersucht werden, entfällt die Beschränkung auf die 10 ähnlichsten Vektoren. Aussortiert wird hierbei jegliche Flexion des Ausgangsterms (Beispielsweise entfällt der gleichartige Term „Banken“ für den Term „Bank“), Abkürzungen (wie „abs“ für „Absatz“) und die bereits erwähnten fälschlich trainierten Terme, also zusammenhanglose Buchstabenkombinationen.

Die Abdeckung der aus C resultierenden Gleichartigen ist analog zu den Ergebnissen aus Kapitel 3.4.1 für GermaNet besser als für Wiktionary. GermaNet deckt sogar 95.03% der Gleichartigen ab, während Wiktionary nur für 68.94% der Gleichartigen einen Eintrag hat.
Beim Vergleich der absoluten Anzahl geht jedoch hervor, dass GermaNet in den meisten Fällen mehr in Relation stehende Gleichartige aufweist als Wiktionary.

Im Wiktionary-Wortnetz stehen prozentual zu den abgedeckten Gleichartigen mehr Terme in Relation zum Ausgangsterm. Für die Terme, die in Wiktionary einen Eintrag haben, korreliert die Vernetzung also besonders gut mit der Vektorähnlichkeit im *word embedding*. Betrachtet man aber, welches Wortnetz summiert die meisten Gleichartigen eines Terms in einer Relation zu diesem kodiert, ist GermaNet ein besser vernetzter Kandidat. Dies zeigt einmal mehr den vielleicht größten Nachteil Wiktionarys gegenüber GermaNet auf: den Umfang.
Zwar sind viele Wiktionary-Artikel umfangreich und stellen syntaktische, semantische und lexikalische Beziehungen zwischen Termen ausführlich dar, jedoch fehlt es an Vollständigkeit. Dabei fallen sowohl unfertige Artikel, in denen die Liste der in Relation stehenden Wörtern noch unvollständig ist, sowie gänzlich fehlende Artikel ins Gewicht.

Ein besonders gutes Beispiel ist hier der Term „Bank“. Im *word embedding* wurden 12 Terme mit einer Vektorähnlichkeit von mehr als 0.6 identifiziert. Von diesen 12 Termen sind in GermaNet 9 eingetragen und 8 stehen dazu in Relation zu dem Term „Bank“. Im symmetrischen Wiktionary-Wortnetz hingegen existieren nur für 6 der Gleichartigen Einträge, jedoch steht jeder dieser Einträge in einer Relation zu „Bank“.
Dabei sind die 6 Terme im Wiktionary-Wortnetz eine Untermenge der 8 Terme aus GermaNet. Lediglich „Geschäftsbank“ und „Kreditbank“ sind nicht im Wiktionary-Wortnetz enthalten und stehen somit auch nicht in Relation zu „Bank“. „Kreditbank“ wurde in Wiktionary allerdings bereits als Unterbegriff zu „Bank“ vermerkt. Zusätzlich dazu ist dort der Begriff „Bankengruppe“ als Wortbildung notiert, der sich zwar auch im *word embedding* als Gleichartiger qualifiziert, jedoch von keinem der beiden Wortnetze abgedeckt wird (Wortbildungen wurden nicht geparst).

Wieder zeigt sich also das große Potenzial von Wiktionary. Häufig besteht in den Artikeln sogar bereits das Grundgerüst zu den Relationen, die Einträge fehlen jedoch noch. Steigt die Arbeit, die freiwillige Nutzer in das Projekt stecken, wird auch das resultierende Wortnetz GermaNet in vielerlei Hinsicht Vorteile abringen.

3.5 Vergleich über eine Umfrage

Der letzte Vergleich nutzt schließlich die Anwender der deutschen Sprache selbst, um zu bestimmen, ob die in den Wortnetzen konstruierten Zusammenhänge das tatsächlich Sprachbild gut repräsentieren. Über eine Online-Umfrage wurden Teilnehmer gebeten, Ober-, Unterbegriffe und Synonyme zu besonders ambiguen Wörtern der deutschen Sprache aufzuzählen (ein Screenshot der Umfrage findet sich im Anhang, Abbildung 11). Dabei wurde die Antonymität bewusst außen vor gelassen. Nach einer eigenständigen Pilot-Umfrage wurde festgestellt, dass durchschnittliche Muttersprachler für diese Relation deutlich weniger Intuition aufzeigten und es somit schwer viel, passende Begriffe zu finden. Für die anderen drei Relationen bestand bereits ein besseres Gefühl, welches dann zum identifizieren geeigneter Begriffe genutzt werden konnte. Schwer taten sich die Befragten auch mit Nicht-Substantiven (besonders bei Ober- und Unterbegriffen), weshalb die Umfrage auf Substantive beschränkt wurde.
Als Orientierungshilfe wurden den Fragebögen außerdem Tests hinzugefügt, die EuroWordNet [Vos02], ein internationales Wortnetz, zur Bestimmung einer zwischen zwei Termen bestehenden Relation nutzt. Den Teilnehmern wurde jedoch nicht aufgetragen, Entscheidungen von diesen Tests abhängig zu machen.

Der Fragebogen wurde in verschiedenen Foren und sozialen Netzwerken veröffentlicht. Um Daten für mehr Begriffe zu bekommen, als einem einzelnen Teilnehmer über einen Fragebogen zugemutet werden sollte, wurden insgesamt vier gleiche Fragebögen erstellt, die sich nur durch die abgefragten Begriffe unterscheiden. Die Fragebögen wurden dann so verteilt, dass sie jeweils einzigartig in bestimmten Foren oder Netzwerken auftauchen, so dass ein Teilnehmer nicht zwei Fragebögen ausfüllt.
Für die insgesamt 20 Begriffe (siehe Anhang) aus der Umfrage wurden so über 1 300 singuläre Terme gesammelt, die zu einem dieser Begriffe in einer Relation stehen. Zusätzlich zur einfachen Abdeckungsuntersuchung wird die Mehfrachnennung von Begriffen als Gewicht für die Relevanz der Relation zu eben diesen Begriffen gewertet werden. Wird beispielsweise „Gebäude“ besonders oft als Synonym zu „Haus“ genannt, wird es besonders bestraft, wenn der Term „Gebäude“ in einem Wortnetz nicht als Synonym zu „Haus“ eingetragen ist.

3.5.1 Abdeckung der Resultate

Zunächst wird wieder die Abdeckung der Wörter in den beiden Netzen und wie viele der in den Umfragen zusammengekommenen Begriffe tatsächlich in einer Relation stehen untersucht. Die Terme, die von den Teilnehmern genannt wurden, werden als **Resultate** bezeichnet, während die Begriffe, zu denen die Teilnehmer passende, in Relation stehende Terme gesucht haben, als **Proben** bezeichnet werden.
Die Abdeckung der Resultate ist in beiden Wortnetzen hoch. Der übliche Unterschied zwischen GermaNet mit einer Abdeckung von 74.08% und Wiktionary mit einer Abdeckung von 64.03% der Resultate besteht auch hier. Nicht abgedeckt sind überwiegend Mehrwortlexeme (wie „Art und Weise“, welches lediglich im Wiktionary-Wortnetz fehlt, da es als Wortverbindung eingetragen ist und somit nicht geparst wurde) und Flexionen (zum Beispiel der Plural: „Vögel“).

Ungewichtet	Wiktionary (N)	Wiktionary (S)	GermaNet
⌀% Synonyme	*10.51%*	*13.98%*	*8.97%*
⌀% Unterbegriffe	*12.83%*	*17.50%*	*17.94%*
⌀% Oberbegriffe	*5.04%*	*7.19%*	*8.95%*

Tabelle 8: Durchschnittliche prozentuale Anzahl an Resultaten, die im jeweiligen Wortnetz in Relation zu der entsprechenden Probe stehen.

Gewichtet	Wiktionary (N)	Wiktionary (S)	GermaNet
⌀% Synonyme	*21.28%*	*26.76%*	*15.64%*
⌀% Unterbegriffe	*16.56%*	*23.24%*	*22.36%*
⌀% Oberbegriffe	*9.27%*	*12.07%*	*12.47%*

Tabelle 9: Durchschnittliche prozentuale Anzahl an Resultaten, die im jeweiligen Wortnetz in Relation zu der entsprechenden Probe stehen. Hierbei wurde das mehrfach Zählen eines unter den Resultaten vorkommenden Terms erlaubt.

Bei der Betrachtung der Anzahl an Resultaten die in Relation zu den Proben stehen, dargestellt in Tabelle 8, fällt zunächst auf, dass schon im nativen Wiktionary-Wortnetz mehr der von den Teilnehmern genannten Synonyme mit den kodierten Synonymen übereinstimmen. Dieses Ergebnis polarisiert sich zusätzlich für den symmetrischen Fall. In den Ober- und Unterbegriffen jedoch decken sich die Resultate weniger mit der Wiktionary kodierung.
Grund hierfür könnte die deutlich stringentere Synonymbeziehung in GermaNet sein. In einem Synset befinden sich durchnittlich weniger als 1.3 Lexeme. Selbst bei hochgradig ambiguen Termen wie „Flügel“, welches sich in GermaNet in sieben Synsets befindet, kommen dabei nur verhältnismäßig wenige Synonyme zusammen: „Schwinge“, „Parteiflügel“ und „Rotorblatt“.
Ein weiteres Beispiel ist „Haus“, welches im Wiktionary-Wortnetz doppelt so viele Synonyme besitzt wie in GermaNet. Darunter sind allerdings auch Terme wie „Wohnhaus“ oder „Wohngebäude“, welche in GermaNet als Unter- beziehungsweise Oberbegriff zu Haus kodiert sind.

Interessant ist auch der große Unterschied zwischen Ober- und Unterbegriffen in allen Wortnetzen. Dieser kommt allerdings daher, dass Oberbegriffe den Teilnehmern besonders schwervielen. So findet sich beispielsweise unter den Resultaten der Term „Huhn“, welcher ein Oberbegriff zu „Flügel“ sein soll. Auch ist die Anzahl der genannten Oberbegriffe selten wesentlich kleiner als die Anzahl der Unterbegriffe. In etwa für jeden zweiten Unterbegriff wurde auch ein Oberbegriff genannt.
Um solchen Fehlern entgegenzuwirken werden im Folgenden die Resultate nach ihrer Häufigkeit gewichten.

3.5.2 Gewichtete Resultate

Anstatt das bloße Vorkommen in einer Relation zu untersuchen, wird nun zusätzlich die Häufigkeit eines Resultats genutzt, um dieses zu gewichten. Informell werden also

Resultate, die häufig genannt wurden, härter bestraft, wenn sie nicht in entsprechender Relation stehen, als Resultate, die nur selten genannt wurden.
Hierfür wird wieder (wie in den Ergebnissen aus Tabelle 8) der prozentuale Anteil der Resultate geprüft, die in Relation stehen, gezählt wird aber diesmal jedes Resultat so oft, wie es von den Teilnehmern genannt wurde.
Angenommen zu „Raum“ seien zwei mal „Küche“ und einmal „Zimmer“ als Unterbegriffe genannt worden und in einem Wortnetz wäre lediglich „Zimmer“ als Unterbegriff kodiert. Dann wären im ungewichteten Fall 50% der Resultate als Unterbegriff kodiert. Im gewichteten Fall jedoch, da der doppelt genannte Term „Küche“ fehlt, wäre die Abdeckung der Resultate in der Unterbegriffsbeziehung nur 33%, da von den drei genannten Termen nur einer in als Unterbegriff vom Wortnetz kodiert wird.

In Tabelle 9 finden sich die angepassten Ergebnisse. Wie erwartet ergeben sich in jedem Fall bessere Ergebnisse, da schlecht eingeschätzte Begriffe seltener genannt werden und somit nicht so stark ins Gewicht fallen.
Besonders Wiktionary profitiert von der Gewichtung. Für Synonyme fällt das Ergebnis mit etwa 27% fast doppelt so gut aus. Dies verstärkt zusätzlich den Unterschied zu GermaNet, dessen Ergebnis sich nicht ganz so stark verbessert.
Der durch die Gewichtung gewonnene Zuwachs ist pro Relationstyp in allen Wortnetzen etwa ähnlich. In allen Fällen ist der Zuwachs bei den Synonymen am größten und am kleinsten bei den Oberbegriffen. Es zeigt sich, wie gut die Intuition der Teilnehmer zu den jeweiligen Relationstypen ist. Der besonders starke Zuwachs der Synonyme deutet auf einen hohen Unterschied zwischen Resultaten, die häufig genannt wurden und Resultaten, die selten oder gar nur einmal genannt wurden hin.
Für „Raum“ wurde beispielsweise 17 mal „Zimmer“ als Synonym genannt, alle anderen 25 Synonyme im Schnitt nur 1,5 mal. Da „Zimmer“ im Wiktionary-Wortnetz als Synonym zu „Raum“ eingetragen ist, in GermaNet jedoch nicht, ergibt sich ein besonders großer Vorteil im Wiktionary-Wortnetz.

Insgesamt lässt sich hier ein positives Fazit für Wiktionary ziehen. Der Vergleich mit gewichteten Resultaten zeigt, dass besonders Synonyme im Wiktionary-Wortnetz umfangreicher eingetragen sind. Damit repräsentiert es das Empfinden der deutschen Muttersprachler eher als GermaNet.
Aber auch im Bezug auf Ober- und Unterbegriffe sind die Ergebnisse spätestens im symmetrischen Wiktionary-Wortnetz gleichauf.
Es lässt sich erahnen, dass Wiktionary trotz der bereits festgestellten Probleme mit Ober- und Unterbegriffsbeziehungen die semantische Nähe zweier Terme ähnlich gut darstellt wie GermaNet. Unabhängig von den Relationstypen sind im symmetrischen Wiktionary-Wortnetz insgesamt mehr Resultate in einer Relation eingetragen. Terme, die eine gewissen semantische Ähnlichkeit aufweisen, liegen also auch in Wiktionary ähnlich nah beisammen wie in GermaNet. In welcher Relation ein bestimmter Term dann jedoch eingetragen ist unterscheidet sich stärker, da Wiktionary-Artikel nicht die Stringenz eines GermaNet-Eintrags aufweisen.
Dies ist jedoch, abhängig davon, wofür das Wortnetz genutzt wird, kein entscheidender Nachteil.

4 Fazit

Zuletzt werden alle Vergleiche und deren Ergebnisse noch einmal kurz veranschaulicht und schließlich ein Fazit zu den Qualitäten Wiktionarys als Wortnetz gezogen. Dabei wird auch auf die Grenzen dieser Arbeit eingegangen und weitreichendere Forschungsmöglichkeiten aufgezeigt.

4.1 Resultate und Evaluation

Nachfolgend werden stichpunktartig die Vergleichskriterien und insbesondere deren Ergebnisse beschrieben. Dies dient einer überschaubaren Betrachtung des umfangreichen Vergleichs, dessen Details im vorangegangenen Kapitel dargelegt wurden. Anschließend werden die Ergebnisse evaluiert.

- **Untersuchung der quantitativen Statistiken**:
  Wiktionary umfasst 30% mehr Synsets als GermaNet und kodiert damit mehr Lexeme als GermaNet.
- **Statistische Untersuchung der Relationen**:
  Verben und Adjektive sind in GermaNet deutlich besser und damit wesentlich umfangreicher in Ober- und Unterbegriffsbeziehungen strukturiert, Wiktionary kodiert dafür wesentlich mehr Antonymrelationen.
- **Statistische Untersuchung der Bedeutungen**:
  Wiktionary teilt die Bedeutungsunterschiede ambiguer Wörter granularer auf, Terme sind in mehr Bedeutungen aufgeteilt.
- **Qualitative Untersuchung der Vernetzung**:
  Relationen in Wiktionary sind asymmetrisch eingetragen. Ober- und Unterbegriffsbeziehungen sind zusätzlich von inkonsistenter Struktur.
- **Statistische Abdeckungsuntersuchung**:
  GermaNet umfasst 25% mehr Terme als Wiktionary, die Schnittmenge der beiden Wortnetze ist jedoch verhältnismäßig klein. Ein Großteil der Terme aus Wiktionary ist einzigartig.
- **Untersuchung der Wiktionary-Domänen**:
  Terme aus Wiktionary sind eher spezieller, Terme in GermaNet eher allgemeiner. Alemannische Dialekte sind in Wiktionary besser vertreten.
- **Abdeckungsuntersuchung des Gut1 Grundwortschatz 500**:
  Erweitertes Parsing ermöglicht Wiktionary auch das Kodieren von anderen Wortarten. GermaNet bleibt auf Substantive, Verben und Adjektive beschränkt.
- **Abdeckungsuntersuchung des Projekts deutscher Wortschatz**:
  GermaNet ist zwar umfangreicher, der Unterschied zu Wiktionary wird jedoch geringer.

- **Abdeckungsuntersuchung von Neologismen aus verschiedenen Quellen**:
  Wiktionary nimmt aktuelle Wörter schneller auf und deckt allgemein aktuellere Begriffe besser ab als GermaNet.

- **Vergleich über ein word embedding**:
  Wiktionary kodiert syntaktische und semantische Nähe besser als GermaNet, leidet jedoch unter dem geringeren Umfang.

- **Vergleich über eine Umfrage**:
  Im Wiktionary-Wortnetz finden sich in der Umgebung eines Terms (in seinen Relationen) ähnliche Terme wie in GermaNet, jedoch teilweise in unterschiedlichen Relationen.

  Wie vor- oder nachteilhaft die festgestellten Eigenschaften des Wiktionary-Wortnetzes sind hängt von dem jeweiligen Anwendungsgebiet ab. Dennoch wird ein möglichst allgemeines Fazit gezogen.

  Der größte und eindeutigste Nachteil Wiktionarys ist der geringe Umfang. Während das aus Wiktionary resultierende Wortnetz mit inkonsistenten und asymmetrischen Relationen kämpft lassen sich zusätzlich 20% weniger Terme nachschlagen als im wohlstrukturierten, stringenten GermaNet. Besonders problematisch ist dies bei Verben und Adjektiven, die aufgrund fehlender künstlicher Konzepte dramatisch schlechter vernetzt sind als in GermaNet. Schließlich wird dieses Problem zusätzlich dadurch polarisiert, dass in dieser Arbeit eine GermaNet Version aus dem Jahr 2014 mit einem Wiktionary Dump von 2017 verglichen wird.

  Dennoch lässt sich aus dem deutschen Wiktionary ein brauchbares Wortnetz konstruieren, das sich mit bestimmten Qualitäten auszeichnet. Größter Vorteil gegenüber den von Linguisten aufgebauten Wortnetzen wie GermaNet ist wohl die Aktualität und die Abdeckung modernster Begriffe. Hier lässt sich für Wiktionary ein Trumpf beim Einsatz in Sozialen Netzwerken vermuten, da die sich schnell entwickelnden Neologismen zügig ins Wortnetz übernommen werden.
  Für den alemannischen Sprachraum könnte sich Wiktionary ebenso als besonders vorteilhafter Kandidat eignen, da es im direkten Vergleich mit GermaNet einen deutlich höheren Anteil an Termen aus den alemannischen Dialekten aufzeigte.

  Auch die Vernetzung des Wiktionary-Wortnetzes steht, abseits der bereits erwähnten Probleme mit den Relationen, GermaNet nicht nach. Für die meisten Terme besteht sogar eine feinere Aufteilung in die verschiedenen Bedeutungen. Vergleiche mit anderen Strukturen wie *word embeddings* haben gezeigt, dass das Wiktionary-Wortnetz die syntaktische und semantische Nähe der Terme zusätzlich zu den lexikalischen Zusammenhängen stärker abbildet als GermaNet.
  Über das Parsen weiterer, vorhandener Kategorien (wie zum Beispiel Wortbildungen, charakteristische Wortkombinationen, Sprichwörter, Redewendungen oder sinnverwandte Wörter) lässt sich dies womöglich noch weiter ausbauen. Insgesamt enthält das deutsche Wiktionary noch einiges mehr an Informationen, die im Sinne des Vergleichs nicht extrahiert wurden und ein aus dem Wörterbuch hervorgehendes Wortnetz bereichern könnten.

  Insgesamt ist das Wiktionary-Wortnetz im direkten Vergleich zu GermaNet noch eine zumeist schlechtere Wahl. Jedoch wächst Wiktionary zügig und wird praktisch

täglich zu einem potenziell besseren Wortnetz. Mit der Zeit könnten sich die Nachteile minimieren und Wiktionary zu einer universell soliden Alternative gegenüber anderen Wortnetzen werden.

4.2 Zukünftige Arbeit

Wie bereits mehrfach erwähnt, ist das Wortnetz, das für diese Arbeit aus Wiktionary gebaut wurde, zu Vergleichszwecken an einigen Stellen begrenzt. Es empfiehlt sich also auch die restlichen, verfügbaren Informationen aus Wiktionary zu extrahieren und deren Einfluss auf das resultierende Wortnetz zu untersuchen.
Auch lässt sich das Problem der asymmetrischen Relationen noch weiter behandeln. Über *word-sense disambiguation* ließe sich beispielsweise eine asymmetrische Relation auf die passende Bedeutung zurückführen und somit die ungenauen Relationen vermeiden. Auch die automatisierte Korrektur inkonsistenter Hypo- beziehungsweise Hyperonymie ist dadurch möglicherweise realisierbar. Somit lässt sich mutmaßlich eine dem GermaNet deutlich ähnlichere, synsetbasierte, Struktur konstruieren, die weniger an den in dieser Arbeit beschriebenen qualitativen Vernetzungsproblemen leidet.

Eine zusätzliche Untersuchungsmöglichkeit wäre außerdem, das Wiktionary-Wortnetz in verschiedenen Anwendungsbereichen zu testen. Auch der Vergleich mit anderen Wortnetzen bietet sich an, wie beispielsweise dem deutschen OpenThesaurus[18], welches das automatisches Generieren und die manuelle Pflege eines Wortnetzes kombiniert [Nab05].

Das ständige Wachstum des Wiktionarys allein mag Anlass zu weiterer Forschung sein. Ähnliche Arbeiten mit dem englischen Wiktionary, welches über neun mal so groß ist wie das deutsche, zeigten bereits, dass Wiktionary das Potenzial zu einem konkurrenzfähigen Wortnetz hat [MG12]. Deutsche Wortnetze sind noch nicht so umfangreich wie englische, jedoch wird auch dieser Unterschied mit der Zeit geringer und das deutsche Wiktionary zu einem ähnlich umfangreichen Wörterbuch heranwachsen.

[18] `https://www.openthesaurus.de/`

A Anhang

Beispiele für Gleichartige aus dem Word2vec embedding

Die Auswahl der händisch zu untersuchten Terme viel auf folgende Begriffe:

Haus, Seite, Flügel, Zug, Satz, Krone, Verbindung, Bank, Story, Schnitt, Bild, Grad, Bruch, Zugang, Land, Figur, Punkt, Raum, Ring, Schlag, Ordnung, Wirbel, Band, Gatter, Kopf, Tour, Absatz, Programm, Welle, Lauf, halten, aufnehmen, kommen, schließen, spielen, zeigen, gehen, nehmen, setzen, schlagen, hart, frei, schwarz, scharf, fein, offen, dünn, klar, fest, dick

Im Folgenden vier Beipsiele, zu denen Gleichartige nach den Kriterien aus Kapitel 3.4.2 im Word2vec embedding identifiziert und um falsche Terme bereinigt wurden. Die Ergebnisse sind in JSON Syntax und zur Veranschaulichung zusätzlich strukturiert.

- "Haus# Substantiv": ["nachbarhaus", "gartenhaus", "häuschen", "wohnhaus", "landhaus", "stadthaus", "sommerhaus", "anwesen", "familienhaus", "patrizierhaus", "schlösschen", "vorderhaus", "stadtpalais", "nebenhaus", "haupthaus", "privathaus", "holzhaus", "eckhaus", "wochenendhaus", "gartenhäuschen", "jagdhaus", "bürgerhaus", "grundstück", "rückgebäude", "zimmer", "doppelhaus", "arbeitszimmer", "hinterhaus", "burghaus", "turmhaus", "stammhaus", "fachwerkhaus", "atelierhaus"]
- "Wirbel# Substantiv": ["wirbelkörper", "schwanzwirbel", "schädelknochen", "lendenwirbel", "rückenwirbel", "brustwirbel", "zahnreihen", "fortsätze", "nasenöffnungen", "zähne", "halswirbel", "hautlappen", "schwanzflosse", "schultergürtel", "orbita", "tentakel", "oberkiefer", "seitenzähne", "augenhöhle", "schuppenreihen", "brustflossen"]
- "zeigen# Verb": ["weisen", "belegen", "aufweisen", "bezeugen", "verdeutlichen", "deuten", "veranschaulichen", "darstellen", "kennzeichnen", "andeuten", "beinhalten", "aufweist", "erkennen"]
- "schwarz# Adjektiv": ["blau", "rot", "grün", "weiß", "grau", "gelb", "neongelb", "braun", "jaecki", "skasa", "dunkelblau", "königsblau"]

Beispiel für ein Ergebnis aus der Umfrage

Die Auswahl der im Fragebogen abgefragten Terme viel auf folgende Begriffe:

Absatz, Bank, Figur, Flügel, Haus, Krone, Lauf, Programm, Punkt, Raum, Ring, Satz, Schlag, Schnitt, Seite, Story, Tour, Verbindung, Welle, Zug

Hier beispielhaft die Ergebnisse aus der Umfrage (siehe Kapitel 3.5) für den Term „Raum“. Jedem Ergebnis ist seine Häufigkeit zugewiesen. Das Beispiel ist in JSON Syntax und zur Veranschaulichung zusätzlich strukturiert.

"Raum": {

- "hyperonyms": { "all": 1, "definition": 1, "dimension": 2, "entfaltungsmöglichkeiten": 1, "gebäude": 2, "geometrie": 1, "haus": 4, "lebensraum": 1, "maßeinheit": 1, "menge": 1, "ort": 1, "orte": 1, "philosophische": 1, "teil des gebäudes": 1, "universum": 2, "wohnung": 4, "wohnungsbereich": 1, "zimmer": 1 },
- "hyponyms": { "abstellplatz": 1, "abstellraum": 3, "ankleidezimmer": 1, "arbeitszimmer": 1, "aufenthaltsort": 1, "bad": 4, "badezimmer": 2, "büro": 2, "ecke": 1, "esszimmer": 4, "fahrradkeller": 1, "flur": 2, "garage": 1, "gastraum": 1, "haus": 1, "hobbykeller": 1, "hotelzimmer": 1, "kammer": 1, "keller": 3, "kinderzimmer": 4, "küche": 6, "ladenlokal": 1, "möbel": 1, "röhrenraum": 1, "schlafraum": 1, "schlafzimmer": 8, "toilette": 1, "unterraum": 1, "vakuumraum": 1, "vektorraum": 1, "wand": 1, "waschküche": 2, "waschraum": 1, "weltraum": 2, "werkkeller": 1, "wohnraum": 4, "wohnung": 1, "wohnzimmer": 7, "zimmer": 3 },
- "synonyms": { "all": 2, "dacht": 1, "eingegrenztes": 1, "freiheit": 2, "freiraum": 2, "gebiet": 1, "gebäudeteil": 1, "hörsaal": 1, "kammer": 3, "kreis": 1, "körper": 1, "landkreis": 1, "ort": 3, "platz": 5, "rückzugsraum": 1, "sphäre": 1, "stube": 1, "umgebung": 1, "vier ecken": 1, "volumen": 3, "weltraum": 1, "wohnung": 1, "wohnungsbereich": 1, "wohnzimmer": 1, "zeit": 2, "zimmer": 17 }

}

Abbildung 11: Benutzerschnittstelle des Fragebogens.

Literatur

[BE99] BARZILAY, Regina ; ELHADAD, Michael: Using Lexical Chains for Text Summarization. In: *Advances in automatic text summarization* (1999), S. 111–121

[BH06] BUDANITSKY, Alexander ; HIRST, Graeme: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. In: *Computational Linguistics* 32 (2006), Nr. 1, S. 13–47

[BP02] BANERJEE, Satanjeev ; PEDERSEN, Ted: An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In: *International Conference on Intelligent Text Processing and Computational Linguistics* Springer, 2002, S. 136–145

[CEE+09] CARSTENSEN, Kai-Uwe ; EBERT, Christian ; EBERT, Cornelia ; JEKAT, Susanne ; LANGER, Hagen ; KLABUNDE, Ralf: *Computerlinguistik und Sprachtechnologie: Eine Einführung*. Springer-Verlag, 2009

[GCPT07] GLOBERSON, Amir ; CHECHIK, Gal ; PEREIRA, Fernando ; TISHBY, Naftali: Euclidean Embedding of Co-occurrence Data. In: *Journal of Machine Learning Research* 8 (2007), Nr. Oct, S. 2265–2295

[GEQ12] GOLDHAHN, Dirk ; ECKART, Thomas ; QUASTHOFF, Uwe: Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: *LREC*, 2012, S. 759–765

[HF97] HAMP, Birgit ; FELDWEG, Helmut: GermaNet-a Lexical-Semantic Net for German. In: *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997, S. 9–15

[KL94] KNIGHT, Kevin ; LUK, Steve K.: Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: *AAAI* Bd. 94, 1994, S. 773–778

[LC13] LEBRET, Rémi ; COLLOBERT, Ronan: Word Embeddings through Hellinger PCA. In: *arXiv preprint arXiv:1312.5542* (2013)

[MCCD13] MIKOLOV, Tomas ; CHEN, Kai ; CORRADO, Greg ; DEAN, Jeffrey: Efficient Estimation of Word Representations in Vector Space. In: *arXiv preprint arXiv:1301.3781* (2013)

[MG12] MEYER, Christian M. ; GUREVYCH, Iryna: Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In: GRANGER, Sylviane (Hrsg.) ; PAQUOT, Magali (Hrsg.): *Electronic Lexicography*. Oxford University Press, November 2012, Kapitel 13, S. 259–291. – ISBN 978–0–19–965486–4

[Mil95] MILLER, George A.: WordNet: a lexical database for English. In: *Communications of the ACM* 38 (1995), Nr. 11, S. 39–41

[MSC+13] MIKOLOV, Tomas ; SUTSKEVER, Ilya ; CHEN, Kai ; CORRADO, Greg S. ; DEAN, Jeff: Distributed Representations of Words and Phrases and their Compositionality. In: *Advances in neural information processing systems*, 2013, S. 3111–3119

[Nab05] NABER, Daniel: OpenThesaurus: ein offenes deutsches Wortnetz. In: *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung, Bonn, Germany* (2005), S. 422–433

[SM98] SCOTT, Sam ; MATWIN, Stan: Text Classification Using WordNet Hypernyms. In: *Use of WordNet in natural language processing systems: Proceedings of the conference*, 1998, S. 38–44

[Vos02] VOSSEN, PJTM: *EuroWordNet: General Document*. Technical report, University of Amsterdam, July 2002

Abbildungsverzeichnis

1 GermaNet Konzepte . . . 3
2 Unterschiede im Grad der Polysemie pro Term . . . 11
3 Symmetrisch versus asymmetrisch eingetragene Relationen im Wiktionary-Wortnetz . . . 12
4 Inkonsistente Hyponymrelation . . . 13
5 Differenz und Schnitt der im Wiktionary-Wortnetz und GermaNet eingetragenen Terme. . . . 14
6 Domänen mit mehr als 50 im Wiktionary-Wortnetz zugeordneten Bedeutungen . . . 15
7 Prozentuale Abdeckung der mit Domänen gekennzeichneten Terme aus dem Wiktionary in GermaNet . . . 16
8 Abdeckung der PdW Terme in Wiktionary und GermaNet pro Jahr . . . 18
9 Differenz der prozentualen Abdeckung der PdW Terme im Wiktionary-Wortnetz und GermaNet pro Jahr . . . 19
10 Abdeckung der Wortwarte Terme in Wiktionary und GermaNet pro Jahr . 21
11 Benutzerschnittstelle des Fragebogens. . . . 34

Tabellenverzeichnis

1 Anzahl an Synsets in Wiktionary und GermaNet . . . 8
2 Relationen in GermaNet . . . 9
3 Relationen in Wiktionary (nativ) . . . 9
4 Relationen in Wiktionary (symmetrisch) . . . 10
5 Terme mit inkonsistenten Hypo-/ Hyperonymrelationen im nativen und symmetrischen Wiktionary-Wortnetz . . . 13
6 Abdeckung der Gleichartigen zu Termen aus S im Wiktionary-Wortnetz und GermaNet. . . . 23
7 Abdeckung der in Relation stehenden abgedeckten Gleichartigen zu Termen aus T_W und T_G im Wiktionary-Wortnetz und GermaNet. . . . 24
8 Durchschnittliche prozentuale Anzahl an Resultaten, die im jeweiligen Wortnetz in Relation zu der entsprechenden Probe stehen. . . . 27
9 (Gewichtete) Durchschnittliche prozentuale Anzahl an Resultaten, die im jeweiligen Wortnetz in Relation zu der entsprechenden Probe stehen. . . . 27